2023 FRM®
Exam Prep



SchweserNotes™
Quantitative Analysis

PART I BOOK 2



KAPLAN® SCHWESER

VISIT…

LANZAROTE Caliente.com

Book 2: Quantitative Analysis

SchweserNotes™ 2023

FRM Part I

KAPLAN SCHWESER

SCHWESERNOTES™ 2023 FRM® PART I BOOK 2: QUANTITATIVE ANALYSIS

©2023 Kaplan, Inc. All rights reserved.

10 9 8 7 6 5 4 3 2 1

ISBN: 978-1-0788-3108-6

Required Disclaimer: GARP® does not endorse, promote, review, or warrant the accuracy of the products or services offered by Kaplan Schweser of FRM® related information, nor does it endorse any pass rates claimed by the provider. Further, GARP® is not responsible for any fees or costs paid by the user to Kaplan Schweser, nor is GARP® responsible for any fees or costs of any person or entity providing any services to Kaplan Schweser. FRM®, GARP®, and Global Association of Risk Professionals™ are trademarks owned by the Global Association of Risk Professionals, Inc.

These materials may not be copied without written permission from the author. The unauthorized duplication of these notes is a violation of global copyright laws. Your assistance in pursuing potential violators of this law is greatly appreciated.

Disclaimer: The SchweserNotes should be used in conjunction with the original readings as set forth by GARP®. The information contained in these books is based on the original readings and is believed to be accurate. However, their accuracy cannot be guaranteed nor is any warranty conveyed as to your ultimate exam success.

CONTENTS

Readings and Learning Objectives

STUDY SESSION 4—PROBABILITY AND STATISTICS

READING 12

Fundamentals of Probability

Exam Focus
Module 12.1: Basics of Probability
Module 12.2: Conditional, Unconditional, and Joint Probabilities
Key Concepts
Answer Key For Module Quizzes

READING 13

Random Variables

Exam Focus
Module 13.1: Probability Mass Functions, Cumulative Distribution Functions, and Expected Values
Module 13.2: Mean, Variance, Skewness, and Kurtosis
Module 13.3: Probability Density Functions, Quantiles, and Linear Transformations
Key Concepts
Answer Key for Module Quizzes

READING 14

Common Univariate Random Variables

Exam Focus
Module 14.1: Uniform, Bernoulli, Binomial, and Poisson Distributions
Module 14.2: Normal and Lognormal Distributions
Module 14.3: Student's *t*, Chi-Squared, and *F*-Distributions
Key Concepts
Answer Key for Module Quizzes

READING 15

Multivariate Random Variables

Exam Focus
Module 15.1: Marginal and Conditional Distributions for Bivariate Distributions
Module 15.2: Moments of Bivariate Random Distributions
Module 15.3: Behavior of Moments for Bivariate Random Variables
Module 15.4: Independent and Identically Distributed Random Variables
Key Concepts

Answer Key for Module Quizzes

STUDY SESSION 5—SAMPLE MOMENTS AND HYPOTHESIS TESTING

READING 16
Sample Moments

Exam Focus
Module 16.1: Estimating Mean, Variance, and Standard Deviation
Module 16.2: Estimating Moments of the Distribution
Key Concepts
Answer Key for Module Quizzes

READING 17
Hypothesis Testing

Exam Focus
Module 17.1: Hypothesis Testing Basics
Module 17.2: Hypothesis Testing Results
Key Concepts
Answer Key for Module Quizzes

STUDY SESSION 6—REGRESSION ANALYSIS

READING 18
Linear Regression

Exam Focus
Module 18.1: Regression Analysis
Module 18.2: Ordinary Least Squares Estimation
Module 18.3: Hypothesis Testing
Key Concepts
Answer Key for Module Quizzes

READING 19
Regression with Multiple Explanatory Variables

Exam Focus
Module 19.1: Multiple Regression
Module 19.2: Measures of Fit in Linear Regression
Key Concepts
Answer Key for Module Quizzes

READING 20
Regression Diagnostics

Exam Focus
Module 20.1: Heteroskedasticity and Multicollinearity
Module 20.2: Model Specification
Key Concepts

Answer Key for Module Quizzes

STUDY SESSION 7—FORECASTING, CORRELATION, AND MACHINE LEARNING

READING 21
Stationary Time Series

Exam Focus
Module 21.1: Covariance Stationary
Module 21.2: Autoregressive and Moving Average Models
Module 21.3: Autoregressive Moving Average (ARMA) Models
Key Concepts
Answer Key for Module Quizzes

READING 22
Non-Stationary Time Series

Exam Focus
Module 22.1: Time Trends
Module 22.2: Seasonality
Module 22.3: Unit Roots
Key Concepts
Answer Key for Module Quizzes

READING 23
Measuring Return, Volatility, and Correlation

Exam Focus
Module 23.1: Defining Returns and Volatility
Module 23.2: Normal and Nonnormal Distributions
Module 23.3: Correlations and Dependence
Key Concepts
Answer Key for Module Quizzes

READING 24
Simulation and Bootstrapping

Exam Focus
Module 24.1: Monte Carlo Simulation and Sampling Error Reduction
Module 24.2: Bootstrapping and Random Number Generation
Key Concepts
Answer Key for Module Quizzes

READING 25
Machine Learning Models

Exam Focus
Module 25.1: Machine Learning and Data Preparation
Module 25.2: Principal Components Analysis and K-Means Clustering
Module 25.3: Methods of Prediction and Sample Splitting

Module 25.4: Reinforcement Learning and Natural Language Processing
Key Concepts
Answer Key for Module Quizzes

READING 26

Machine Learning and Prediction

Exam Focus
Module 26.1: Categorial Variables, Regularization, and Logistic Regression
Module 26.2: Decision Trees, Ensemble Learning, K-Nearest Neighbors, and Support Vector Machines
Module 26.3: Neural Networks and Performance
Key Concepts
Answer Key for Module Quizzes

Formulas
Appendix
Index

READINGS AND LEARNING OBJECTIVES

STUDY SESSION 4

12. Fundamentals of Probability

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 1.

After completing this reading, you should be able to:
a. describe an event and an event space.
b. describe independent events and mutually exclusive events.
c. explain the difference between independent events and conditionally independent events.
d. calculate the probability of an event for a discrete probability function.
e. define and calculate a conditional probability.
f. distinguish between conditional and unconditional probabilities.
g. explain and apply Bayes' rule.

13. Random Variables

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 2.

After completing this reading, you should be able to:
a. describe and distinguish a probability mass function from a cumulative distribution function, and explain the relationship between these two.
b. understand and apply the concept of a mathematical expectation of a random variable.
c. describe the four common population moments.
d. explain the differences between a probability mass function and a probability density function.
e. characterize the quantile function and quantile-based estimators.
f. explain the effect of a linear transformation of a random variable on the mean, variance, standard deviation, skewness, kurtosis, median, and interquartile range.

14. Common Univariate Random Variables

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 3.

After completing this reading, you should be able to:
a. distinguish the key properties and identify the common occurrences of the following distributions: uniform distribution, Bernoulli distribution, binomial distribution, Poisson distribution, normal distribution, lognormal distribution, Chi-squared distribution, Student's t and F-distributions.
b. describe a mixture distribution and explain the creation and characteristics of mixture distributions.

15. Multivariate Random Variables

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 4.

After completing this reading, you should be able to:
a. explain how a probability matrix can be used to express a probability mass function.
b. compute the marginal and conditional distributions of a discrete bivariate random variable.
c. explain how the expectation of a function is computed for a bivariate discrete random variable.
d. define covariance and explain what it measures.

e. explain the relationship between the covariance and correlation of two random variables and how these are related to the independence of the two variables.
f. explain the effects of applying linear transformations on the covariance and correlation between two random variables.
g. compute the variance of a weighted sum of two random variables.
h. compute the conditional expectation of a component of a bivariate random variable.
i. describe the features of an independent and identically distributed (iid) sequence of random variables.
j. explain how the iid property is helpful in computing the mean and variance of a sum of iid random variables.

STUDY SESSION 5

16. Sample Moments

Global Association of Risk Professionals. *Quantitative Analysis.* New York, NY: Pearson, 2022. Chapter 5.

After completing this reading, you should be able to:
a. estimate the mean, variance, and standard deviation using sample data.
b. explain the difference between a population moment and a sample moment.
c. distinguish between an estimator and an estimate.
d. describe the bias of an estimator and explain what the bias measures.
e. explain what is meant by the statement that the mean estimator is BLUE.
f. describe the consistency of an estimator and explain the usefulness of this concept.
g. explain how the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) apply to the sample mean.
h. estimate and interpret the skewness and kurtosis of a random variable.
i. use sample data to estimate quantiles, including the median.
j. estimate the mean of two variables and apply the CLT.
k. estimate the covariance and correlation between two random variables.
l. explain how coskewness and cokurtosis are related to skewness and kurtosis.

17. Hypothesis Testing

Global Association of Risk Professionals. *Quantitative Analysis.* New York, NY: Pearson, 2022. Chapter 6.

After completing this reading, you should be able to:
a. construct an appropriate null hypothesis and alternative hypothesis and distinguish between the two.
b. differentiate between a one-sided and a two-sided test and identify when to use each test.
c. explain the difference between Type I and Type II errors and how these relate to the size and power of a test.
d. understand how a hypothesis test and a confidence interval are related.
e. explain what the *p*-value of a hypothesis test measures.
f. construct and apply confidence intervals for one-sided and two-sided hypothesis tests and interpret the results of hypothesis tests with a specific confidence level.
g. identify the steps to test a hypothesis about the difference between two population means.
h. explain the problem of multiple testing and how it can lead to biased results.

STUDY SESSION 6

18. Linear Regression

Global Association of Risk Professionals. *Quantitative Analysis.* New York, NY: Pearson, 2022. Chapter 7.

After completing this reading, you should be able to:

a. describe the models that can be estimated using linear regression and differentiate them from those which cannot.
b. interpret the results of an ordinary least squares (OLS) regression with a single explanatory variable.
c. describe the key assumptions of OLS parameter estimation.
d. characterize the properties of OLS estimators and their sampling distributions.
e. construct, apply, and interpret hypothesis tests and confidence intervals for a single regression coefficient in a regression.
f. explain the steps needed to perform a hypothesis test in a linear regression.
g. describe the relationship among a *t*-statistic, its *p*-value, and a confidence interval.
h. estimate the correlation coefficient from the R^2 measure obtained in linear regressions with a single explanatory variable.

19. Regression with Multiple Explanatory Variables

Global Association of Risk Professionals. *Quantitative Analysis.* New York, NY: Pearson, 2022. Chapter 8.

After completing this reading, you should be able to:
a. distinguish between the relative assumptions of single and multiple regression.
b. interpret regression coefficients in a multiple regression.
c. interpret goodness-of-fit measures for single and multiple regressions, including R^2 and adjusted R^2.
d. construct, apply, and interpret joint hypothesis tests and confidence intervals for multiple coefficients in a regression.
e. calculate the regression R^2 using the three components of the decomposed variation of the dependent variable data: the explained sum of squares, the total sum of squares and the residual sum of squares.

20. Regression Diagnostics

Global Association of Risk Professionals. *Quantitative Analysis.* New York, NY: Pearson, 2022. Chapter 9.

After completing this reading, you should be able to:
a. explain how to test whether a regression is affected by heteroskedasticity.
b. describe approaches to using heteroskedastic data.
c. characterize multicollinearity and its consequences; distinguish between multicollinearity and perfect collinearity.
d. describe the consequences of excluding a relevant explanatory variable from a model and contrast those with the consequences of including an irrelevant regressor.
e. explain two model selection procedures and how these relate to the bias-variance tradeoff.
f. describe the various methods of visualizing residuals and their relative strengths.
g. describe methods for identifying outliers and their impact.
h. determine the conditions under which OLS is the best linear unbiased estimator.

STUDY SESSION 7

21. Stationary Time Series

Global Association of Risk Professionals. *Quantitative Analysis.* New York, NY: Pearson, 2022. Chapter 10.

After completing this reading, you should be able to:
a. describe the requirements for a series to be covariance-stationary.
b. define the autocovariance function and the autocorrelation function.
c. define white noise, and describe independent white noise and normal (Gaussian) white noise.
d. define and describe the properties of autoregressive (AR) processes.
e. define and describe the properties of moving average (MA) processes.
f. explain how a lag operator works.

g. explain mean reversion and calculate a mean-reverting level.
h. define and describe the properties of autoregressive moving average (ARMA) processes.
i. describe the application of AR, MA, and ARMA processes.
j. describe sample autocorrelation and partial autocorrelation.
k. describe the Box-Pierce Q statistic and the Ljung-Box Q statistic.
l. explain how forecasts are generated from ARMA models.
m. describe the role of mean reversion in long-horizon forecasts.
n. explain how seasonality is modeled in a covariance-stationary ARMA.

22. Non-Stationary Time Series

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 11.

After completing this reading, you should be able to:
a. describe linear and nonlinear time trends.
b. explain how to use regression analysis to model seasonality.
c. describe a random walk and a unit root.
d. explain the challenges of modeling time series containing unit roots.
e. describe how to test if a time series contains a unit root.
f. explain how to construct an h-step-ahead point forecast for a time series with seasonality.
g. calculate the estimated trend value and form an interval forecast for a time series.

23. Measuring Return, Volatility, and Correlation

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 12.

After completing this reading, you should be able to:
a. calculate, distinguish, and convert between simple and continuously compounded returns.
b. define and distinguish among volatility, variance rate, and implied volatility.
c. describe how the first two moments may be insufficient to describe non-normal distributions.
d. explain how the Jarque-Bera test is used to determine whether returns are normally distributed.
e. describe the power law and its use for non-normal distributions.
f. define correlation and covariance and differentiate between correlation and dependence.
g. describe properties of correlations between normally distributed variables when using a one-factor model.
h. compare and contrast the different measures of correlation used to assess dependence.

24. Simulation and Bootstrapping

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 13.

After completing this reading, you should be able to:
a. describe the basic steps to conduct a Monte Carlo simulation.
b. describe ways to reduce Monte Carlo sampling error.
c. explain the use of antithetic and control variates in reducing Monte Carlo sampling error.
d. describe the bootstrapping method and its advantage over Monte Carlo simulation.
e. describe pseudo-random number generation.
f. describe situations where the bootstrapping method is ineffective.
g. describe the disadvantages of the simulation approach to financial problem solving.

25. Machine Learning Models

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 14.

After completing this reading, you should be able to:
a. discuss the philosophical and practical differences between machine-learning techniques and classical econometrics.
b. explain the differences among the training, validation, and test data sub-samples, and how each is used.

c. understand the differences between and consequences of underfitting and overfitting, and propose potential remedies for each.
d. use principal components analysis to reduce the dimensionality of a set of features.
e. describe how the K-means algorithm separates a sample into clusters.
f. be aware of natural language processing and how it is used.
g. differentiate among unsupervised, supervised, and reinforcement learning models.
h. explain how reinforcement learning operates and how it is used in decision-making.

26. Machine Learning and Prediction

Global Association of Risk Professionals. *Quantitative Analysis*. New York, NY: Pearson, 2022. Chapter 15.

After completing this reading, you should be able to:
a. explain the role of linear regression and logistic regression in prediction.
b. understand how to encode categorical variables.
c. discuss why regularization is useful, and distinguish between the ridge regression and LASSO approaches.
d. show how a decision tree is constructed and interpreted.
e. describe how ensembles of learners are built.
f. outline the intuition behind the K nearest neighbors and support vector machine methods for classification.
g. understand how neural networks are constructed and how their weights are determined.
h. evaluate the predictive performance of logistic regression models and neural network models using a confusion matrix.

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 1.

READING 12

FUNDAMENTALS OF PROBABILITY

Study Session 4

EXAM FOCUS

This reading covers important terms and concepts associated with probability theory. Specifically, we will examine the difference between independent and mutually exclusive events, discrete probability functions, and the difference between unconditional and conditional probabilities. Bayes' rule is also examined as a way to update a given set of prior probabilities. For the exam, be able to calculate conditional probabilities, joint probabilities, and probabilities based on a probability function. Also, understand when and how to apply Bayes' formula.

MODULE 12.1: BASICS OF PROBABILITY

When an outcome is unknown, such as the outcome (realization) of the flip of a coin or the high temperature tomorrow in Dubai, we refer to it as a **random variable**. We can describe a random variable with the probabilities of its possible outcomes. For the flip of a fair coin, we refer to the probability of heads as P(heads), which is 50%. We can think of a probability as the likelihood that an outcome will occur. If we flip a fair coin 100 times, we expect that on average it will be heads 50 times.

A probability equal to 0 for an outcome means that the outcome will not happen. A probability equal to 1 for an outcome means it will happen with certainty. Probabilities cannot be less than 0 or greater than 1.

The probability that a random variable will have a specific outcome, given that some other outcome has occurred, is referred to as a **conditional probability**. The probability that A will occur, given that B has occurred, is written as P(A | B). For example, the probability that a day's high temperature in Seattle will be between 70 and 80 degrees is an **unconditional probability** (i.e., *marginal probability*). The probability that the high temperature will be between 70 and 80 degrees, given that the sky is cloudy that day, is a conditional probability.

The probability that both A and B will occur is written P(AB) and referred to as the **joint probability** of A and B (both occurring).

Events and Event Spaces

LO 12.a: Describe an event and an event space.

An **event** is a single outcome or a combination of outcomes for a random variable. Consider a random variable that is the result of rolling a fair six-sided die. The outcomes with positive probability (those that may happen) are the integers 1, 2, 3, 4, 5, and 6. For the event x = 3, we can write P(3) = 1/6 = 16.7%. Other possible events include getting a 3 *or* 4, P(3 or 4) = 2/6 = 33.3%, and getting an even number, P(x is even) = P(x = 2, 4, or 6) = 3/6 = 50%. The probability that the realization of this random variable is equal to one of the possible outcomes (x = 1, 2, 3, 4, 5, or 6) is 100%.

The **event space** for a random variable is the set of all possible outcomes and combinations of outcomes. Consider a flip of a fair coin. The event space is heads, tails, heads and tails, and neither heads nor tails. P(heads) and P(tails) are both 50%. The probability of both heads and tails is zero, as is the probability of neither heads nor tails.

PROFESSOR'S NOTE

The notation $P(A \cup B)$ is sometimes used to mean the probability of A *or* B, and the notation $P(A \cap B)$ is sometimes used to mean the probability of A *and* B.

Independent and Mutually Exclusive Events

LO 12.b: Describe independent events and mutually exclusive events.

Two events are **independent events** if knowing the outcome of one does not affect the probability of the other. When two events are independent, the following two probability relationships must hold:

1. P(A) × P(B) = P(AB). The probability that both A and B will happen is the product of their unconditional probabilities.
2. P(A | B) = P(A). The conditional probability of A given that B occurs is simply the unconditional probability of A occurring. This means B occurring does not change the probability of A.

Consider flipping a coin twice. Getting heads on the first flip does not change the probability of getting heads on the second flip. The two events are independent. In this case, the **joint probability** of getting heads on both flips is simply the product of their unconditional expectations. Given that the probability of getting heads is 50%, the probability of getting heads on two flips in a row is 0.5 × 0.5 = 25%.

If A_1 , A_2 , ..., A_n are independent events, their joint probability $P(A_1 \text{ and } A_2 \dots \text{ and } A_n)$ is equal to $P(A_1) \times P(A_2) \times \dots \times P(A_n)$.

Two events are **mutually exclusive events** if they cannot both happen. Consider the possible outcomes of one roll of a die. The events "x = an even number" and "x = 3" are mutually exclusive; they cannot both happen on the same roll.

In general, P(A or B) = P(A) + P(B) − P(AB). We must subtract the probability of both A and B happening to avoid counting those outcomes twice. If the probability that one stock will rise tomorrow, P(A), is 60% and the probability that another stock will rise tomorrow, P(B), is 55%, we cannot calculate the probability that both will rise tomorrow as 60% + 55% = 115%. We must subtract the joint probability that both stocks will rise to get P(A or B).

When events A and B are mutually exclusive, P(AB) is zero, so P(A or B) is simply P(A) + P(B).

Conditionally Independent Events

LO 12.c: Explain the difference between independent events and conditionally independent events.

Two conditional probabilities, P(A | C) and P(B | C), may be independent or dependent regardless of whether the unconditional probabilities, P(A) and P(B), are independent or not. When two events are **conditionally independent events**, P(A | C) × P(B | C) = P(AB | C).

Consider Event A, "scores above average on an exam," and Event B, "is taller than average." For a population of grade school students, these events may not be independent, as taller students are older on average and likely in a higher grade. Taller students may well do better on a given exam than shorter (younger) students. If we add the conditioning Event C "age equals 8," we may find that height and exam scores are independent, that is, P(A | C) and P(B | C) are independent while P(A) and P(B) are not.

MODULE QUIZ 12.1

1. For the roll of a fair six-sided die, how many of the following are classified as events?
   - The outcome is 3.
   - The outcome is an even number.
   - The outcome is not 2, 3, 4, 5, or 6.

   A. One.
   B. Two.
   C. Three.
   D. None.

2. Which of the following equalities does not imply that the events A and B are independent?
   A. P(AB) = P(A) × P(B).
   B. P(A or B) = P(A) + P(B) – P(AB).
   C. P(A | B) = P(A).
   D. P(AB) / P(B) = P(A).

3. Two independent events:
   A. must be conditionally independent.
   B. cannot be conditionally independent.
   C. may be conditionally independent or not conditionally independent.
   D. are conditionally independent only if they are mutually exclusive.

MODULE 12.2: CONDITIONAL, UNCONDITIONAL, AND JOINT PROBABILITIES

Discrete Probability Function

LO 12.d: Calculate the probability of an event for a discrete probability function.

A **discrete probability function** is one for which there are a finite number of possible outcomes. The probability function gives us the probability of each possible outcome. Consider a random variable for which the possible outcomes are x = 1, 2, 3, or 4, with a probability function of x/10 so that P(x) = x/10. The probability of an outcome of 3 is 3/10 = 30%. The probability of an outcome of either 2 or 4 is 2/10 + 4/10 = 60%. This function qualifies as a probability function because the probability of getting one of the possible outcomes is 1/10 + 2/10 + 3/10 + 4/10 = 10/10 = 100%.

Conditional and Unconditional Probabilities

LO 12.e: Define and calculate a conditional probability.

LO 12.f: Distinguish between conditional and unconditional probabilities.

Sometimes we are interested in the probability of an event, given that some other event has occurred. As mentioned earlier, we refer to this as a **conditional probability**, P(A | B).

Consider conditional probabilities that an employee at Acme, Inc., earns more than $40,000 per year, P(40+), conditioned on the highest level of education an employee has attained. Employees fall into one of three education levels: no degree (ND), bachelor's degree (BD), and higher-than-bachelor's degree (HBD). If 60% of the employees have no degree, 30% of the employees have attained only a bachelor's degree, and 10% have attained a higher degree, we write P(ND) = 60%, P(BD) = 30%, and P(HBD) = 10%.

Note that the three levels of education attainment are *mutually exclusive*; an employee can only be in one of the three categories of educational attainment. Note also that the three categories are also *exhaustive*; the categories cover all the possible levels of educational attainment. We can write this as P(ND or BD or HBD) = 100%.

Given a conditional probability and the unconditional probability of the conditioning event, we can calculate the **joint probability** of both events using P(AB) = P(A | B) × P(B). Assume that for Acme, 10% of the employees with no degree, 70% of the employees with only a bachelor's degree, and 100% of employees with a degree beyond a bachelor's degree earn more than $40,000 per year. That is, P(40+ | ND) = 10%, P(40+ | BD) = 70%, and P(40+ | HBD) = 100%.

Using these conditional probabilities, along with the unconditional probabilities P(ND) = 60%, P(BD) = 30%, and P(HBD) = 10%, we can calculate the joint probabilities:

P(40+ and ND) = 10% × 60% = 6%

$P(40+ \text{ and } BD) = 70\% \times 30\% = 21\%$

$P(40+ \text{ and } HBD) = 100\% \times 10\% = 10\%.$

We can use these probabilities to illustrate the **total probability rule**, which states that if the conditioning events B_i are mutually exclusive and exhaustive then:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_n)P(B_n)$$

This is the sum of the joint probabilities. For Acme, we have P(40+) = 6% + 21% + 10% = 37% of the employees earn more than $40,000 per year.

Rearranging $P(AB) = P(A \mid B) \times P(B)$, we get:

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

That is, we can calculate a conditional probability from the joint probability of two events and the unconditional probability of the conditioning event. As an example, the conditional probability is $P(40+ \mid BD)$ is:

$$\frac{P(40 + \text{and } BD)}{P(BD)} = \frac{21\%}{30\%} = 70\%$$

Bayes' Rule

LO 12.g: Explain and apply Bayes' rule.

Bayes' rule allows us to use information about the outcome of one event to improve our estimates of the unconditional probability of another event.

From our rules of probability, we know that $P(A \mid B) \times P(B) = P(AB)$ and that $P(B \mid A) \times P(A) = P(AB)$, so we can write $P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$. Rearranging these terms, we can arrive at Bayes' rule:

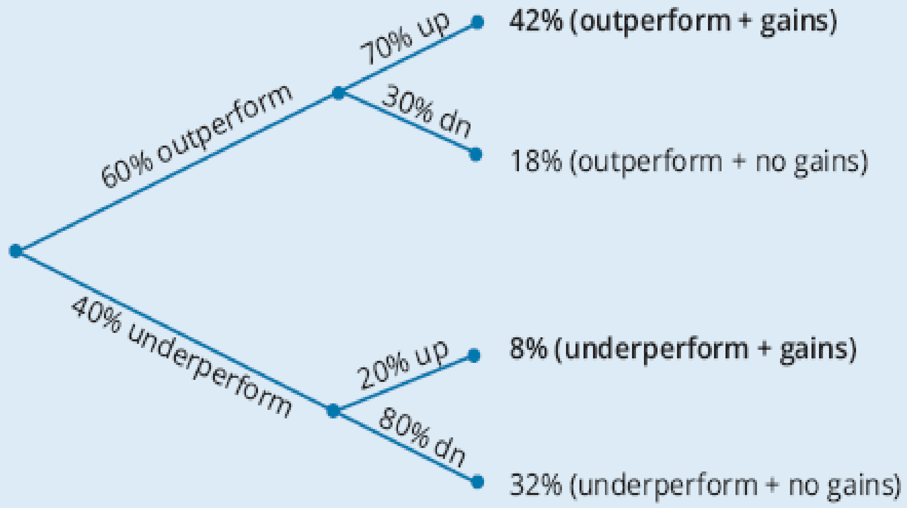
$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

Given the unconditional probabilities of A and B and the conditional probability of B given A, we can calculate the conditional probability of A given B. The following example illustrates the use of Bayes' rule and provides some intuition about what this formula is telling us.

EXAMPLE: Bayes' formula

There is a 60% probability the economy will outperform, and if it does, there is a 70% probability a stock will go up and a 30% probability the stock will go down. There is a 40% probability the economy will underperform, and if it does, there is a 20% probability the stock in question will increase in value (have gains) and an 80% probability it will not. Given that the stock increased in value, **calculate** the probability that the economy outperformed.

Answer:



In the earlier figure, we have multiplied the probabilities to calculate the probabilities of each of the four outcome pairs. Note that these sum to 1. Given that the stock has gains, what is our updated probability of an outperforming economy? We sum the probability of stock gains in both states (outperform and underperform) to get 42% + 8% = 50%. Given that the stock has gains, the probability that the economy has outperformed is:

$$\frac{42\%}{50\%} = 84\%.$$

The numerator for the calculation of the updated probability P(A | B) using Bayes' formula in the example is the joint probability of outperform and gains. This is calculated as P(gains | outperform) × P(outperform) (i.e., 0.7 × 0.6 = 0.42). The denominator is the unconditional probability of gains, P(gains | outperform) + P(gains | underperform) (i.e., 0.42 + 0.08 = 0.50).

EXAMPLE: **Probability concepts and relationships**

A shipment of 1,000 cars has been unloaded into a parking area. The cars have the following features:

- There are 600 blue (B) cars.
- Of the blue cars, 150 have driver assist (DA) technology.
- There are 400 red (R) cars.
- Of the red cars, 200 have DA technology.

Given these facts, **calculate** the following:

1. Unconditional probabilities: P(B) and P(R)
2. Conditional probabilities: P(DA | B) and P(DA | R)
3. Joint probabilities: P(B and DA) and P(R and DA)
4. Total probability rule: P(DA)
5. Bayes' rule: P(B | DA)

Answer:

Unconditional probabilities:

P(B) = 600/1,000 = 60%

P(R) = 400/1,000 = 40%

Conditional probabilities:

P(DA | B) = 150/600 = 25%

P(DA | R) = 200/400 = 50%

Joint probabilities:

P(B and DA) = P(DA | B)P(B) = 25%(60%) = 15%; 15%(1,000) = 150 of the cars are blue with driver assist

P(R and DA) = P(DA | R)P(R) = 50%(40%) = 20%; 20%(1,000) = 200 of the cars are red with driver assist

Total probability rule:

P(DA) = P(DA | B)P(B) + P(DA | R)P(R) = 25%(60%) + 50%(40%) = 35%; 35% (1,000) = 350 of the cars have driver assist

Bayes' rule:

P(B | DA) = P(B and DA)/P(DA) = 15%/35% = 42.9%; 350 cars have driver assist and of those cars, 150 are blue: 150/350 = 0.42857 = 42.9%

Independence:

Now, assume we add to our information that 40% of the blue cars (240) are convertibles and 40% of the red cars (160) are convertibles, so that 400 of the cars are convertibles. In this case, P(B | C) = 240/400 = 60% = P(B) and P(R | C) = 160/400 = 40% = P(R). This meets the requirement for independence that P(A | B) = P(A). The fact that a car chosen at random is a convertible gives us no additional information about whether a car is blue or red.

MODULE QUIZ 12.2

1. The probability function for the outcome of one roll of a six-sided die is given as P(X) = x/21. What is P(x > 4)?
   A. 16.6%.
   B. 23.8%.
   C. 33.3%.
   D. 52.4%.

2. The relationship between the probability that both Event A and Event B will occur and the conditional probability of Event A given that Event B occurs is:

A. $P(AB) = P(A|B)P(B)$.

B. $P(A) = \frac{P(A|B)}{P(AB)}$.

C. $P(A) = \frac{P(AB)}{P(A|B)}$.

D. $P(AB) = P(A|B)P(A)$.

3. The probability that shares of Acme will increase in value over the next month is 50% and the probability that shares of Acme and shares of Best will both increase in value over the next month is 40%. The probability that Best shares will increase in value, given that Acme shares increase in value over the next month, is closest to:
   A. 20%.
   B. 40%.
   C. 80%.
   D. 90%.

KEY CONCEPTS

LO 12.a

An event is one of the possible outcomes or a subset of the possible outcomes of a random event, such as the flip of a coin. The event space is all the subsets of possible outcomes and the empty set (none of the possible outcomes).

LO 12.b

Two events are independent if either of the following conditions hold:

- $P(A) \times P(B) = P(AB)$
- $P(A | B) = P(A)$

Two events are mutually exclusive if the joint probability, $P(AB) = 0$ (i.e., both cannot occur). When two events are mutually exclusive, $P(A \text{ or } B) = P(A) + P(B)$.

LO 12.c

If two events conditional on a third event are independent, we say they are conditionally independent. For example, if $P(AB | C) = P(A | C)\ P(B | C)$, then A and B are conditionally independent. Two events may be independent but conditionally dependent, or vice versa.

LO 12.d

A probability function describes the probability for each possible outcome for a discrete probability distribution. For example, $P(x) = x/25$, defined over the outcomes {1, 2, 3, 4, 5}.

LO 12.e

The joint probability of two events, P(AB), is the probability that they will both occur: $P(AB) = P(A | B) \times P(B)$. This relationship can be rearranged to define the conditional probability of A given B as follows:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

LO 12.f

An unconditional probability (i.e., marginal probability) is the probability of an event occurring.

A conditional probability, $P(A \mid B)$, is the probability of an Event A occurring given that Event B has occurred.

LO 12.g

Bayes' rule is:

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

This formula allows us to update the unconditional probability, $P(A)$, based on the fact that B has occurred. $P(AB)$ can be calculated as $P(B \mid A)P(A)$.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 12.1

1. **C** All of the outcomes and combinations specified are included in the event space for the random variable. (LO 12.a)
2. **B** $P(A \text{ or } B) = P(A) + P(B) - P(AB)$ holds for both independent and dependent events. The other equalities are only true for independent events. (LO 12.b)
3. **C** Two independent events may be conditionally independent or not conditionally independent. (LO 12.c)

Module Quiz 12.2

1. **D** The probability of $x > 4$ is the probability of an outcome of 5 or 6 ($5/21 + 6/21 = 52.4\%$). (LO 12.d)
2. **A** The (joint) probability that both A and B will occur is equal to the conditional probability of Event A given that Event B has occurred, multiplied by the unconditional probability of Event B. (LO 12.e)
3. **C** Bayes' formula tells us that:

   $$P(A \mid B) = \frac{P(AB)}{P(B)}$$

   Applying that to the information given, we can write:

   $$P(\text{Best increases} \mid \text{Acme increases}) = \frac{P(\text{Best increases and Acme increases})}{P(\text{Acme increases})}$$

   $40\%/50\% = 80\%$ (LO 12.g)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 2.

READING 13

RANDOM VARIABLES

Study Session 4

EXAM FOCUS

This reading addresses the concepts of expected value, variance, skewness, and kurtosis. The characteristics and calculations of these measures will be discussed. For the exam, be able to distinguish among a probability mass function, a cumulative distribution function, and a probability density function. Also, be able to compute expected value, and be able to identify the four common population moments of a statistical distribution.

MODULE 13.1: PROBABILITY MASS FUNCTIONS, CUMULATIVE DISTRIBUTION FUNCTIONS, AND EXPECTED VALUES

Random Variables and Probability Functions

LO 13.a: Describe and distinguish a probability mass function from a cumulative distribution function, and explain the relationship between these two.

A **discrete random variable** is one that can take on only a countable number of possible outcomes. It can take on only two possible values, zero and one, and is referred to as a Bernoulli random variable. We can model the outcome of a coin flip as a Bernoulli random variable where heads = 1 and tails = 0. The number of days in June that will have a temperature greater than 70 degrees is also a discrete random variable. The possible outcomes are the integers from 0 to 30.

A **continuous random variable** has an uncountable number of possible outcomes. The amount of rainfall that will fall in June is an example of a continuous random variable. There are an infinite number of possible outcomes because for any two values (e.g., 6.95 inches and 6.94 inches), we can find a number between them [e.g., (6.95 + 6.94) / 2 = 6.945]. Because there are an infinite number of possible outcomes, the probability of any single value is zero. For continuous random variables, we measure probabilities only over some positive interval, (e.g., the probability that rainfall in June will be between 6.94 and 6.95 inches).

A **probability mass function (PMF)**, $f(x) = P(X = x)$, gives us the probability that the outcome of a discrete random variable, X, will be equal to a given number, x. For a Bernoulli random variable for which the $P(x = 1) = p$, the PMF is $f(x) = p^x (1 - p)^{1-x}$. This yields $P(x = 1) = p$ and $P(x = 0) = 1 - p$.

A second example of a PMF is $f(x) = 1/6$, which is the probability that one roll of a six-sided die will take on one of the possible outcomes one through six. Each of the possible outcomes has the same probability of occurring (1/6 = 16.67%).

A third example is the PMF $f(x) = x/10$ for a random variable that can take on values of 1, 2, 3, or 4. For example, $P(x = 3) = f(3) = 3/10 = 30\%$.

For all of these PMFs, the sum of the probabilities of all of the possible outcomes is 100%, a requirement for a PMF.

A **cumulative distribution function (CDF)** gives us the probability that a random variable will take on a value less than or equal to x [i.e., $F(x) = P(X \leq x)$].

For a Bernoulli random variable with possible outcomes of zero and one, the CDF is:

$$F(x) = \begin{matrix} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{matrix}$$

While the PMF for this Bernoulli variable is defined only for X = 0 or 1, the corresponding CDF is defined for all real numbers. For example, $P(X < 0.1456) = F(0.1456) = 1 - p$.

For the roll of a six-sided die, the CDF is $F(x) = x/6$, so that the probability of a roll of 3 or less is $F(3) = 3/6 = 50$. This illustrates an important relationship between a PMF and its corresponding CDF; the probability of an outcome less than or equal to x is simply the sum of the probabilities of all the possible outcomes less than or equal to x. For the roll of a six-sided die. $F(3) = f(1) + f(2) + f(3) = 1/6 + 1/6 + 1/6 = 3/6 = 50\%$.

Expectations

LO 13.b: Understand and apply the concept of a mathematical expectation of a random variable.

The **expected value** is the weighted average of the possible outcomes of a random variable, where the weights are the probabilities that the outcomes will occur. The mathematical representation for the expected value of random variable X is:

$$E(X) = \Sigma P(x_i)x_i = P(x_1)x_1 + P(x_2)x_2 + \ldots + P(x_n)x_n$$

Here, E is referred to as the expectations operator and is used to indicate the computation of a probability-weighted average. The symbol x_1 represents the first observed value (observation) for random variable X; x_2 is the second observation, and so on through the nth observation. The concept of expected value may be demonstrated using probabilities associated with a coin toss. On the flip of one coin, the occurrence of the event “heads” may be used to assign the value of one to a random variable.

Alternatively, the event “tails” means the random variable equals zero. Statistically, we would formally write the following:

if heads, then X = 1

if tails, then X = 0

For a fair coin, P(heads) = P(X = 1) = 0.5, and P(tails) = P(X = 0) = 0.5. The expected value can be computed as follows:

$$E(X) = \Sigma P(x_i)x_i = P(X = 0)(0) + P(X = 1)(1) = (0.5)(0) + (0.5)(1) = 0.5$$

In any individual flip of a coin, *X* cannot assume a value of 0.5. Over the long term, however, the average of all the outcomes is expected to be 0.5. Similarly, the expected value of the roll of a fair die, where X = number that faces up on the die, is determined to be:

$$E(X) = \Sigma P(x_i)x_i = (1/6)(1) + (1/6)(2) + (1/6)(3) + (1/6)(4) + (1/6)(5) + (1/6)(6)$$

$$E(X) = 3.5$$

We can never roll a 3.5 on a die, but over the long term, 3.5 should be the average value of all outcomes.

The expected value is, statistically speaking, our best guess of the outcome of a random variable. While a 3.5 will never appear when a die is rolled, the average amount by which our guess differs from the actual outcomes is minimized when we use the expected value calculated this way.

Note that the probabilities of the outcomes for a coin flip (0 or 1) and the probabilities of the outcomes for the roll of a die are equal for all of the possible outcomes in both cases. When outcomes are equally likely, the expected value is simply the mean (average) of the outcomes:

$$\frac{1+0}{2} = 0.5 \text{ for a coin flip}$$

$$\frac{1+2+3+4+5+6}{6} = 3.5 \text{ for the roll of a die}$$

When we estimate the expected value of a random variable based on *n* observations, we use the mean of the observed values as our estimate of the mean of the underlying probability distribution. In terms of a probability model, we are assuming that the outcomes are equally likely, that is, each has a probability of 1/n. Multiplying each outcome by 1/n and then summing them, produces the same expected value as dividing the sum of the outcomes by *n*.

In other cases, the probabilities of the outcomes are not equal and we calculate the expected value as the weighted sum of the outcomes, where the weights are the probabilities of each outcome. The following example illustrates such a case.

EXAMPLE: **Expected earnings per share (EPS)**

The probability distribution of EPS for Ron’s Stores is given in the following figure. **Calculate** the expected earnings per share.

EPS Probability Distribution

Probability	EPS
10%	£1.80
20%	£1.60
40%	£1.20
30%	£1.00
100%	

Answer:

The expected EPS is simply a weighted average of each possible EPS, where the weights are the probabilities of each possible outcome.

$$E(EPS) = 0.10(1.80) + 0.20(1.60) + 0.40(1.20) + 0.30(1.00) = £1.28$$

The following are two useful properties of expected values:

1. If c is any constant, then:

$$E(cX) = cE(X)$$

2. If X and Y are any random variables, then:

$$E(X + Y) = E(X) + E(Y)$$

MODULE QUIZ 13.1

1. The probability mass function (PMF) for a discrete random variable that can take on the values 1, 2, 3, 4, or 5 is $P(X = x) = x/15$. The value of the cumulative distribution function (CDF) of 4, F(4), is equal to:
   A. 26.7%.
   B. 40.0%.
   C. 66.7%.
   D. 75.0%.

2. An analyst has estimated the following probabilities for gross domestic product growth next year:

   P(4%) = 10%, P(3%) = 30%, P(2%) = 40%, P(1%) = 20%

   Based on these estimates, the expected value of GDP growth next year is:
   A. 2.0%.
   B. 2.3%.
   C. 2.5%.
   D. 2.8%.

MODULE 13.2: MEAN, VARIANCE, SKEWNESS, AND KURTOSIS

LO 13.c: Describe the four common population moments.

The population moments most often used are

- mean;
- variance;
- skewness; and
- kurtosis.

The first moment, the mean of a random variable, is its expected value, E(X), which we discussed previously. The mean can be represented by the Greek letter μ (mu).

The other three moments are **central moments** because the functions involve the random variable minus its mean, $X - \mu$. Subtracting the mean produces functions that are unaffected by the location of the mean. These moments give us information about the shape of a probability distribution around its mean.

PROFESSOR'S NOTE

Since central moments are measured relative to the mean, the first central moment equals zero and is, therefore, not typically used.

The second central moment of a random variable is its **variance**, σ^2. Variance is defined as:

$$\sigma^2 = E\{[X - E(X)]^2\} = E[(X - \mu)^2]$$

Squaring the deviations from the mean ensures that σ^2 is positive. Variance gives us information about how widely dispersed the values of the random variable are around the mean.

We often use the square root of variance, σ, as a measure of dispersion because it has the same units as the random variable. If our distribution is for percentage rates of return, the standard deviation is also measured in terms of percentage returns.

The third central moment of a distribution is:

$$E\{[X - E(X)]^3\} = E[(X - \mu)^3]$$

Skewness, a measure of a distribution's symmetry, is the standardized third moment. We standardize it by dividing it by the standard deviation cubed.

$$\text{skewness} = \frac{E[(X - \mu)^3]}{\sigma^3}$$

Because we both subtract the mean and divide by standard deviation cubed, skewness is unaffected by differences in the mean or in the variance of the random variable. This allows us to compare skewness of two different distributions directly. A distribution with skew = 0 is perfectly symmetric.

The fourth central moment of a distribution is:

$$E\{[X - E(X)]^4\} = E[(X - \mu)^4]$$

Kurtosis is the standardized fourth moment.

$$\text{kurtosis} = \frac{E[(X - \mu)^4]}{\sigma^4}$$

Kurtosis is a measure of the shape of a distribution, in particular the total probability in the tails of the distribution relative to the probability in the rest of the distribution. The higher the kurtosis, the greater the probability in the tails of the distribution. We sometimes refer to distributions with high kurtosis as fat-tailed distributions.

The following figures illustrate the concepts of skewness and kurtosis for a probability distribution.

Figure 13.1: Skewness

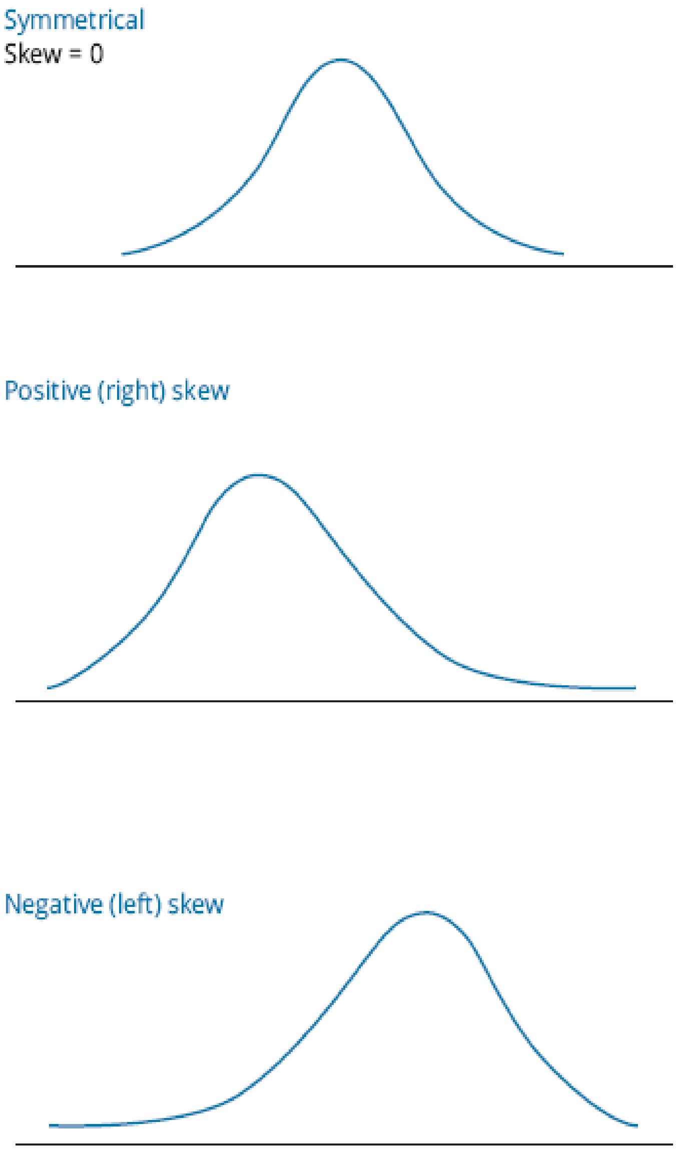
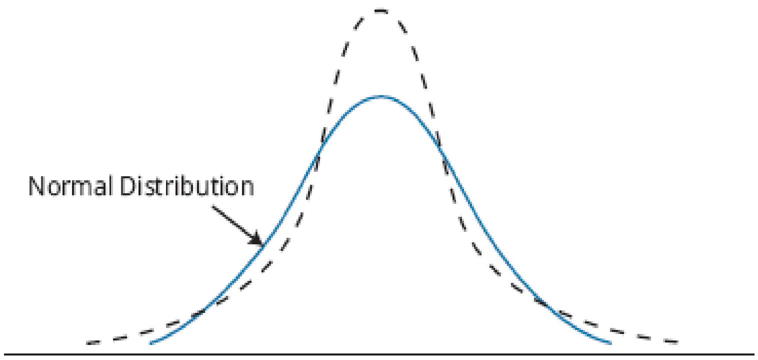


Figure 13.2: Kurtosis



MODULE QUIZ 13.2

1. For two financial securities with distributions of returns that differ only in their kurtosis, the one with the higher kurtosis will have:
   A. a wider dispersion of returns around the mean.
   B. a greater probability of extreme positive and negative returns.
   C. less peaked distribution of returns.
   D. a more uniform distribution.

MODULE 13.3: PROBABILITY DENSITY FUNCTIONS, QUANTILES, AND LINEAR TRANSFORMATIONS

Probability Density Functions

LO 13.d: Explain the differences between a probability mass function and a probability density function.

Recall that we used a PMF to describe the probabilities of the possible outcomes for a discrete random variable. A simple example is P(X = x) = *f* (x) = x/10 for the possible outcomes 1, 2, 3, and 4. The PMF tells us the probability of each of those possible outcomes, P(X = 4) = 4/10 = 40%.

Recall that a continuous random variable can take on any of an infinite number of possible outcomes so that the probability of any single outcome is zero. We describe a continuous distribution function with a **probability density function (PDF)**, rather than a PMF. A PDF allows us to calculate the probability of an outcome between two values (over an interval). This probability is the area under the PDF over the interval. Mathematically, we take the integral of the PDF over an interval to calculate the probability that the random variable will take on a value in that interval.

Quantile Functions

LO 13.e: Characterize the quantile function and quantile-based estimators.

The **quantile function**, Q(*a*), is the inverse of the CDF. Recall that a CDF gives us the probability that a random variable will be less than or equal to some value X = x. The interpretation of the CDF is the same for discrete and continuous random variables.

Consider a CDF that gives us a probability of 30% that a continuous random variable takes on values less than 2 [i.e., P(X < 2) = F(2) = 30%]. The quantile function, Q(30%), for this distribution would return the value 2; 30% of the outcomes are expected to be less than 2. A common use of quantiles is to report the results of standardized tests. Consider a student with a score of 122 on an exam. If the student's quantile score is 74%, this indicates that that the student's score of 122 was higher than 74% of those who took the test. The quantile function, Q(74%), would return the student's score of 122.

Two quantile measures are of particular interest to us here. One is the value of the quantile function for 50%. This is termed the **median** of the distribution. On average,

50% of the variable's outcomes will be below the median and 50% of the variable's outcomes will be above the median. For a symmetric distribution (skew = 0), the mean and median will be equal. For a distribution with positive (right) skew, the median will be less than the mean, but will be greater than the mean for distributions with negative (left) skew.

The second quantile measure of interest here is the **interquartile range (IQR)**. The interquartile range is the upper and lower value of the outcomes of a random variable that include the middle 50% of its probability distribution. The lower value is Q(25%) and the upper value is Q(75%). The lower value is the value that we expect 25% of the outcomes to be less than, and the upper value is the value that we expect 75% of the values to be less than. Like standard deviation, the interquartile range is a measure of the variability of a random variable. Compared to a given distribution, the outcomes of a distribution with a lower interquartile range are more concentrated around the mean, just as they are for a distribution with a lower standard deviation.

Linear Transformations of Random Variables

LO 13.f: Explain the effect of a linear transformation of a random variable on the mean, variance, standard deviation, skewness, kurtosis, median, and interquartile range.

A **linear transformation** of a random variable, X, takes the form Y = a + bX, where *a* and *b* are constants. The constant *a* shifts the location of the random variable, X, and *b* rescales the values of X. The relationships between the moments of the distribution of X and the moments of the distribution of Y, a linear transformation of X, are as follows:

- The mean of Y can be calculated as E(Y) = a + bE(X), both the location and the scale are affected.
- The variance of Y can be calculated as $\sigma_Y^2 = b^2\sigma_X^2$; while *a* shifts the location of the distribution, it does not affect the dispersion around the mean which is rescaled by *b*. The standard deviation of Y is simply $\sigma_Y = \sqrt{b^2\sigma_X^2} = |b|\sigma_X$.
- With *b* > 0 (an increasing transformation), the skew is unaffected, skew Y = skew X.
- With *b* < 0 (a decreasing transformation), the magnitude of the skew is unaffected, but the sign is changed, skew Y = –skew X.
- A linear transformation of X does not affect kurtosis, kurtosis Y = kurtosis X.

MODULE QUIZ 13.3

1. Which of the following regarding a probability density function (PDF) is correct? A PDF:
   A. provides the probability of each of the possible outcomes of a random variable.
   B. can provide the same information as a cumulative distribution function (CDF).
   C. describes the probabilities for any random variable.
   D. only applies to a discrete probability distribution.

2. For the quantile function, Q(x):
   A. the CDF function F[Q(23%)] = 23%.
   B. Q(23%) will identify the largest 23% of all possible outcomes.

C. Q(50%) is the interquartile range.
D. x can only take on integer values.

3. For a random variable, X, the variance of Y = a + bX is:

A. $a^2 + b^2\sigma_x^2$.
B. $b\sigma_x^2$.
C. $b^2\sigma_x^2$.
D. $a + b^2\sigma_x^2$.

KEY CONCEPTS

LO 13.a

A probability mass function (PMF), $f(x)$, gives us the probability that a discrete random variable will take on the value x.

A cumulative distribution function (CDF), F(x), gives us the probability that a random variable X will take on a value less than or equal to x.

LO 13.b

The expected value of a discrete random variable is the probability-weighted average of the possible outcomes (i.e., the mean of the distribution).

LO 13.c

Four commonly used moments of a random variable are its mean, variance (standard deviation), skewness, and kurtosis. The mean is the expected value of the random variable, variance is a measure of dispersion, skewness is a measure of symmetry, and kurtosis is a measure of the proportion of the outcomes in the tails of the distribution.

LO 13.d

A PMF provides the probability that a discrete random variable will take on a given value. A PDF provides the probability that the outcome for a continuous random variable will be within a given interval.

LO 13.e

A quantile is the percentage of outcomes less than a given outcome. A quantile function, Q(x%), provides the value of an outcome which is greater than x% of all possible outcomes. Q(50%) is the median of a distribution. 50% of the outcomes are greater than the median and 50% of the outcomes are less than the median. The interquartile range is an interval that includes the central 50% of all possible outcomes.

LO 13.f

For a variable Y = a + bX (a linear transformation of X):

- the mean of Y is E(Y) = a + bE(X);
- the variance of Y is $\sigma_Y^2 = b^2\sigma_x^2$ and the standard deviation is $\sigma_Y = |b|\sigma_x$;
- the skew of Y = skew X, for b > 0, and skew Y = −skew X for b < 0; and
- the kurtosis of Y = kurtosis X.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 13.1

1. **C** F(4) is the probability that the random variable will take on a value of 4 or less. We can calculate $P(X \leq 4)$ as 1/15 + 2/15 + 3/15 + 4/15 = 66.7%, or by subtracting 5/15, P(X = 5), from 100% to get 66.7%. (LO 13.a)

2. **B** The expected value is computed as: (4)(10%) + (3)(30%) + (2)(40%) + (1)(20%) = 2.3%. (LO 13.b)

Module Quiz 13.2

1. **B** High kurtosis indicates that the probability in the tails (extreme outcomes) are greater (i.e., the distribution will have fatter tails). (LO 13.c)

Module Quiz 13.3

1. **B** A PDF evaluated between minus infinity and a given value gives the probability of an outcome less than the given value; the same information is provided by a CDF. A PDF provides the probabilities only for a continuous random variable. The probability that a continuous random variable will take on a given value is zero. (LO 13.d)

2. **A** Q(23%) gives us a value that is greater than 23% of all outcomes and the CDF for that value is the probability of an outcome less than that value (i.e., 23%). (LO 13.e)

3. **C** The variance of Y is $b^2\sigma_x^2$, where σ_x^2 is the variance of X. (LO 13.f)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 3.

READING 14

COMMON UNIVARIATE RANDOM VARIABLES

Study Session 4

EXAM FOCUS

This reading explores the following common probability distributions: uniform, Bernoulli, binomial, Poisson, normal, lognormal, chi-squared, Student's *t*-, *F*-, exponential, and beta. You will learn the properties, parameters, and common occurrences of these distributions. For the exam, focus most of your attention on the binomial, normal, and Student's *t*-distributions. Also, know how to standardize a normally distributed random variable, how to use a *z*-table, and how to construct confidence intervals.

LO 14.a: Distinguish the key properties and identify the common occurrences of the following distributions: uniform distribution, Bernoulli distribution, binomial distribution, Poisson distribution, normal distribution, lognormal distribution, Chi-squared distribution, Student's t and F-distributions.

MODULE 14.1: UNIFORM, BERNOULLI, BINOMIAL, AND POISSON DISTRIBUTIONS

The Uniform Distribution

The **continuous uniform distribution** is defined over a range that spans between some lower limit, *a*, and some upper limit, *b*, which serve as the parameters of the distribution. Outcomes can only occur between *a* and *b*, and because we are dealing with a continuous distribution, even if $a < x < b$, $P(X = x) = 0$. Formally, the properties of a continuous uniform distribution may be described as follows.

For all $a \leq x_1 < x_2 \leq b$ (i.e., for all x_1 and x_2 between the boundaries *a* and *b*):

- $P(X < a \text{ or } X > b) = 0$ (i.e., the probability of *X* outside the boundaries is zero).
- $P(x_1 \leq X \leq x_2) = (x_2 - x_1) / (b - a)$. This defines the probability of outcomes between x_1 and x_2.

Don't miss how simple this is just because the notation is so mathematical. For a continuous uniform distribution, the probability of outcomes in a range that is one-half the whole range is 50%. The probability of outcomes in a range that is one-quarter of the possible range is 25%.

EXAMPLE: Continuous uniform distribution

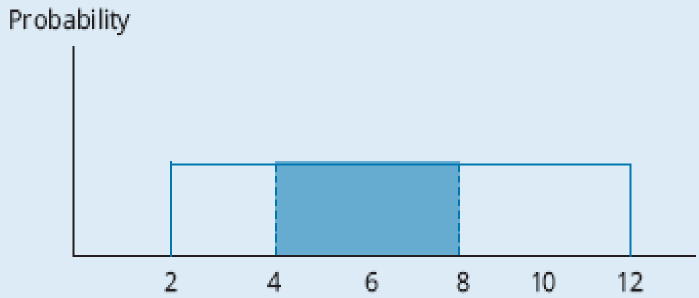
X is uniformly distributed between 2 and 12. **Calculate** the probability that *X* will be between 4 and 8.

Answer:

$$\frac{8-4}{12-2} = \frac{4}{10} = 40\%$$

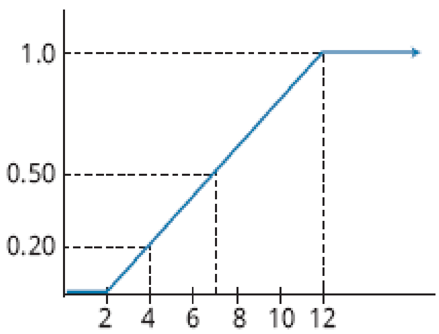
The following figure illustrates this continuous uniform distribution. Note that the area bounded by 4 and 8 is 40% of the total probability between 2 and 12 (which is 100%).

Continuous Uniform Distribution



The **cumulative distribution function (CDF)** is linear over the variable's range. The CDF for the distribution in the previous example, P(X < x), is shown in Figure 14.1.

Figure 14.1: CDF for a Continuous Uniform Variable



The probability density function (PDF) for a continuous uniform distribution is expressed as:

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b, \text{ else } f(x) = 0$$

The mean and variance, respectively, of a uniform distribution are:

$$E(x) = \frac{a+b}{2}$$

$$Var(x) = \frac{(b-a)^2}{12}$$

The Bernoulli Distribution

A **Bernoulli random variable** only has two possible outcomes. The outcomes can be defined as either a *success* or a *failure*. The probability of success, p, may be denoted with the value *1* and the probability of failure, $1 - p$, may be denoted with the value *0*. Bernoulli distributed random variables are commonly used for assessing the probability of binary outcomes, such as the probability that a firm will default on its debt over some interval.

For a Bernoulli random variable for which the $P(x = 1) = p$, the probability mass function is $f(x) = p^x (1 - p)^{1-x}$. This yields $P(x = 1) = p$ and $P(x = 0) = 1 - p$.

For a Bernoulli random variable, $\mu_X = p$ and the variance is given by $Var(X) = p(1 - p)$. Note that the variance is low for values of p close to 1 or 0, and the maximum variance is for $p = 0.5$.

For a Bernoulli random variable with possible outcomes 0 and 1, the CDF is:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

Note that while the probability mass function (PMF) for this Bernoulli variable is defined only for X = 0 or 1, the corresponding CDF is defined for all real numbers.

The Binomial Distribution

A **binomial random variable** may be defined as the number of successes in a given number of Bernoulli trials, whereby the outcome can be either *success* or *failure*. The probability of success, p, is constant for each trial and the trials are independent. Under these conditions, the binomial probability function defines the probability of exactly x successes in n trials. It can be expressed using the following formula:

$$p(x) = P(X = x) = (\text{number of ways to choose } x \text{ from } n)p^x(1-p)^{n-x}$$

where:

$$(\text{number of ways to choose } x \text{ from } n) = \frac{n!}{(n-x)!x!}$$

So, the probability of exactly x successes in n trials is:

$$p(x) = \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}$$

EXAMPLE: Binomial probability

Assuming a binomial distribution, **compute** the probability of drawing three black beans from a bowl of black and white beans if the probability of selecting a black bean in any given attempt is 0.6. You will draw five beans from the bowl.

Answer:

$$P(X = 3) = p(3) = \frac{5!}{2!3!}(0.6)^3(0.4)^2 = (120/12)(0.216)(0.160)$$

$$= 0.3456$$

Some intuition about these results may help you remember the calculations. Consider that a (very large) bowl of black and white beans has 60% black beans and each time you select a bean, you replace it in the bowl before drawing again. We want to know the probability of selecting exactly three black beans in five draws, as in the previous example.

One way this might happen is BBBWW. Because the draws are independent, the probability of this is easy to calculate. The probability of drawing a black bean is 60%, and the probability of drawing a white bean is 1 − 60% = 40%. Therefore, the probability of selecting BBBWW, in order, is 0.6 × 0.6 × 0.6 × 0.4 × 0.4 = 3.456%. This is the $p^3(1 - p)^2$ from the formula and *p* is 60%, the probability of selecting a black bean on any single draw from the bowl. BBBWW is not, however, the only way to choose exactly three black beans in five trials. Another possibility is BBWWB, and a third is BWWBB. Each of these will have exactly the same probability of occurring as our initial outcome, BBBWW. That's why we need to answer the question of how many ways (different orders) there are for us to choose three black beans in five draws. Using the formula, there are $\frac{5!}{(5-3)!3!} = 10$ ways; 10 × 3.456% = 34.56%, the answer we computed previously.

For a given series of *n* trials, the expected number of successes, or E(X), is given by the following formula:

expected value of X = E(X) = np

The intuition is straightforward; if we perform *n* trials and the probability of success on each trial is *p*, we expect *np* successes.

The variance of a binomial random variable is given by:

variance of X = np(1 − p)

EXAMPLE: Expected value of a binomial random variable

Based on empirical data, the probability that the Dow Jones Industrial Average (DJIA) will increase on any given day has been determined to equal 0.67. Assuming the only other outcome is that it decreases, we can state p(UP) = 0.67 and p(DOWN) = 0.33. Further, assume that movements in the DJIA are independent (i.e., an increase in one day is independent of what happened on another day).

Using the information provided, **compute** the expected value of the number of up days in a five-day period.

Answer:

Using binomial terminology, we define success as UP, so p = 0.67. Note that the definition of success is critical to any binomial problem.

$$E(X \mid n = 5, p = 0.67) = (5)(0.67) = 3.35$$

Recall that the " | " symbol means *given*. Hence, the preceding statement is read as: the expected value of *X* given that n = 5, and the probability of success = 67% is 3.35.

Using the equation for the variance of a binomial distribution, we find the variance of *X* to be:

$$Var(X) = np(1 - p) = 5(0.67)(0.33) = 1.106$$

We should note that because the binomial distribution is a discrete distribution, the result X = 3.35 is not possible. However, if we were to record the results of many five-day periods, the average number of up days (successes) would converge to 3.35.

Binomial distributions are used extensively in the investment world where outcomes are typically seen as successes or failures. In general, if the price of a security goes up, it is viewed as a success. If the price of a security goes down, it is a failure. In this context, binomial distributions are often used to create models to aid in the process of asset valuation.

PROFESSOR'S NOTE

We will examine binomial trees for stock option valuation in Book 4.

The Poisson Distribution

The **Poisson distribution** is a discrete probability distribution with a number of real-world applications. For example, the number of defects per batch in a production process or the number of 911 calls per hour are discrete random variables that follow a Poisson distribution.

While the Poisson random variable *X* refers to the *number of successes per unit*, the parameter lambda (λ) refers to the average or *expected number of successes per unit*. The mathematical expression for the Poisson distribution for obtaining *X* successes, given that λ successes are expected, is:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

An interesting feature of the Poisson distribution is that both its mean and variance are equal to the parameter, λ.

EXAMPLE: Using the Poisson distribution (1)

On average, the 911 emergency switchboards receive 0.1 incoming calls per second. Assuming the arrival of calls follows a Poisson distribution, what is the probability

that in a given minute exactly 5.0 phone calls will be received?

Answer:

We first need to convert the seconds into minutes. Note that λ, the expected number of calls per minute, is (0.1)(60) = 6.0. Hence:

$$P(X = 5) = \frac{6^5 e^{-6}}{5!} = 0.1606 = 16.06\%$$

This means that, given the average of 0.1 incoming calls per second, there is a 16.06% chance there will be five incoming phone calls in a minute.

EXAMPLE: Using the Poisson distribution (2)

Assume there is a 0.01 probability of a patient experiencing severe weight loss as a side effect from taking a recently approved drug used to treat heart disease. What is the probability that out of 200 such procedures conducted on different patients, five patients will develop this complication? Assume that the number of patients developing the complication from the procedure is Poisson distributed.

Answer:

Let X = expected number of patients developing the complication from the procedure = np = (200)(0.01) = 2

$$P(X = 5) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{2^5 e^{-2}}{5!} = 0.036 = 3.6\%$$

This means that given a complication rate of 0.01, there is a 3.6% probability that 5 out of every 200 patients will experience severe weight loss from taking the drug.

MODULE QUIZ 14.1

1. If 5% of the cars coming off the assembly line have some defect in them, what is the probability that out of three cars chosen at random, exactly one car will be defective? Assume that the number of defective cars has a Poisson distribution.
   A. 0.129.
   B. 0.135.
   C. 0.151.
   D. 0.174.
2. A recent study indicated that 60% of all businesses have a web page. Assuming a binomial probability distribution, what is the probability that exactly four businesses will have a web page in a random sample of six businesses?
   A. 0.138.
   B. 0.276.
   C. 0.311.
   D. 0.324.
3. What is the probability of an outcome being between 15 and 25 for a random variable that follows a continuous uniform distribution within the range of 12 to 28?
   A. 0.509.
   B. 0.625.
   C. 1.000.
   D. 1.600.

MODULE 14.2: NORMAL AND LOGNORMAL DISTRIBUTIONS

The Normal Distribution

The **normal distribution** is important for many reasons. Many of the random variables that are relevant to finance and other professional disciplines follow a normal distribution. In the area of investment and portfolio management, the normal distribution plays a central role in portfolio theory.

The PDF for the normal distribution is:

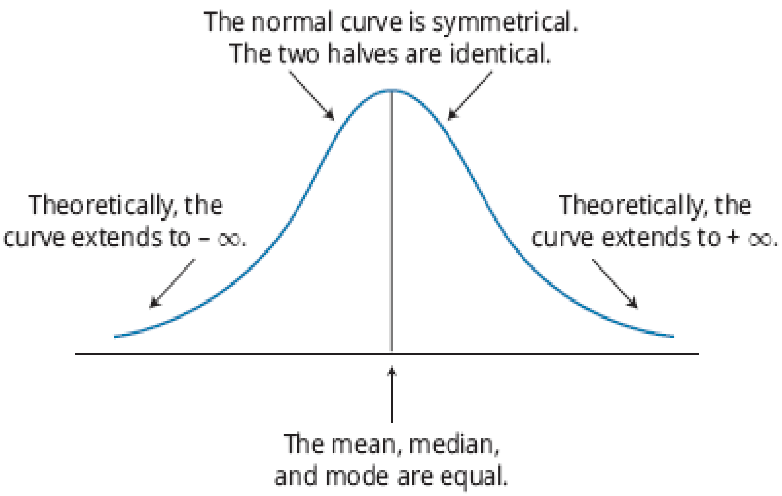
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The normal distribution has the following key properties:

- It is completely described by its mean, μ, and variance, σ^2, stated as $X \sim N(\mu, \sigma^2)$. In words, this says, "X is normally distributed with mean μ and variance σ^2."
- Skewness = 0, meaning the normal distribution is symmetric about its mean, so that $P(X \leq \mu) = P(\mu \leq X) = 0.5$, and mean = median = mode.
- Kurtosis = 3; this is a measure of how the distribution is spread out with an emphasis on the tails of the distribution. Excess kurtosis is measured relative to 3, the kurtosis of the normal distribution.
- A linear combination of normally distributed independent random variables is also normally distributed.
- The probabilities of outcomes further above and below the mean get smaller and smaller but do not go to zero (the tails get very thin but extend infinitely).

Many of these properties are evident from examining the graph of a normal distribution's PDF as illustrated in Figure 14.2.

Figure 14.2: Normal Distribution PDF

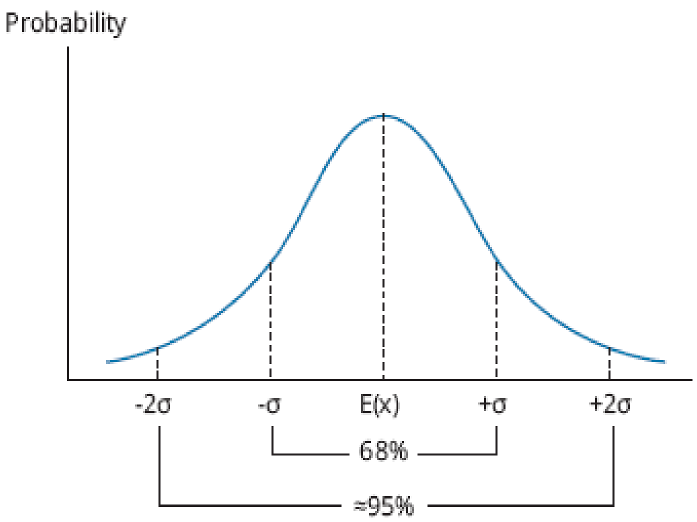


A **confidence interval** is a range of values around the expected outcome within which we expect the actual outcome to be some specified percentage of the time. A 95%

confidence interval is a range that we expect the random variable to be in 95% of the time. For a normal distribution, this interval is based on the expected value (sometimes called a point estimate) of the random variable and on its variability, which we measure with standard deviation.

Confidence intervals for a normal distribution are illustrated in Figure 14.3. For any normally distributed random variable, 68% of the outcomes are within one standard deviation of the expected value (mean), and approximately 95% of the outcomes are within two standard deviations of the expected value.

Figure 14.3: Confidence Intervals for a Normal Distribution



In practice, we will not know the actual values for the mean and standard deviation of the distribution, but will have estimated them as $\bar{X}$ and s. The three confidence intervals of most interest are given by the following:

- The 90% confidence interval for X is $\bar{X} - 1.65s$ to $\bar{X} + 1.65s$.
- The 95% confidence interval for X is $\bar{X} - 1.96s$ to $\bar{X} + 1.96s$.
- The 99% confidence interval for X is $\bar{X} - 2.58s$ to $\bar{X} + 2.58s$.

EXAMPLE: Confidence intervals

The average return of a mutual fund is 10.5% per year and the standard deviation of annual returns is 18%. If returns are approximately normal, what is the 95% confidence interval for the mutual fund return next year?

Answer:

Here μ and σ are 10.5% and 18%, respectively. Thus, the 95% confidence interval for the return, *R*, is:

$$10.5 \pm 1.96(18) = -24.78\% \text{ to } 45.78\%$$

Symbolically, this result can be expressed as:

$$P(-24.78 < R < 45.78) = 0.95 \text{ or } 95\%$$

The interpretation is that the annual return is expected to be within this interval 95% of the time, or 95 out of 100 years.

The standard normal distribution

A standard normal distribution (i.e., *z*-distribution) is a normal distribution that has been standardized so it has a mean of zero and a standard deviation of 1 [i.e., N~(0,1)]. To standardize an observation from a given normal distribution, the *z-value* of the observation must be calculated. The *z*-value represents the number of standard deviations a given observation is from the population mean. *Standardization* is the process of converting an observed value for a random variable to its *z*-value. The following formula is used to standardize a random variable:

$$z = \frac{\text{observation} - \text{population mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

PROFESSOR'S NOTE

The term *z-value* will be used for a standardized observation in this reading. The terms *z-score* and *z-statistic* are also commonly used.

EXAMPLE: Standardizing a random variable (calculating *z*-values)

Assume the annual earnings per share (EPS) for a population of firms are normally distributed with a mean of $6 and a standard deviation of $2.

What are the *z*-values for EPS of $2 and $8?

Answer:

If EPS = x = $8, then z = (x − μ) / σ = ($8 − $6) / $2 = +1

If EPS = x = $2, then z = (x − μ) / σ = ($2 − $6) / $2 = −2

Here, z = +1 indicates that an EPS of $8 is one standard deviation above the mean, and z = −2 means that an EPS of $2 is two standard deviations below the mean.

Calculating probabilities using z-values

Now we will show how to use standardized values (*z*-values) and a table of probabilities for *Z* to determine probabilities. A portion of a table of the CDF for a standard normal distribution is shown in Figure 14.4. We will refer to this table as the *z*-table, as it contains values generated using the cumulative density function for a standard normal distribution, denoted by F(Z). Thus, the values in the *z*-table are the probabilities of observing a *z*-value that is less than a given value, *z* [i.e., P(Z < z)]. The numbers in the first column are *z*-values that have only one decimal place. The columns to the right supply probabilities for *z*-values with two decimal places.

Note that the *z*-table in Figure 14.4 only provides probabilities for positive *z*-values. This is not a problem because we know from the symmetry of the standard normal distribution that F(−Z) = 1 − F(Z). The tables in the back of many texts provide probabilities for negative *z*-values, but we will work with only the positive portion of the table because this may be all you get on the exam. In Figure 14.4, we can find the probability that a standard normal random variable will be less than 1.66, for example. The table value is 95.15%. The probability that the random variable will be less than −1.66 is simply 1 − 0.9515 = 0.0485 = 4.85%, which is also the probability that the variable will be greater than +1.66.

Figure 14.4: Cumulative Probabilities for a Standard Normal Distribution

CDF Values for the Standard Normal Distribution: The z-Table										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.5	.6915	Please note that several of the rows have been deleted to save space.*								
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

*A complete cumulative standard normal table is included in the Appendix.

PROFESSOR'S NOTE

When you use the standard normal probabilities, you have formulated the problem in terms of standard deviations from the mean. Consider a security with returns that are approximately normal, an expected return of 10%, and standard deviation of returns of 12%. The probability of returns greater than 30% is calculated based on the number of standard deviations that 30% is above the expected return of 10%. In this case, 30% is 20% above the expected return of 10%, which is 20 / 12 = 1.67 standard deviations above the mean. We look up the probability of returns less than 1.67 standard deviations above the mean (0.9525 or 95.25% from Figure 14.4) and calculate the probability of returns more than 1.67 standard deviations above the mean as 1 − 0.9525 = 4.75%.

EXAMPLE: Using the *z*-table (1)

Considering again EPS distributed with μ = $6 and σ = $2, what is the probability that EPS will be $9.70 or more?

Answer:

Here we want to know P(EPS > $9.70), which is the area under the curve to the right of the *z*-value corresponding to EPS = $9.70 (see the distribution that follows).

The *z*-value for EPS = $9.70 is:

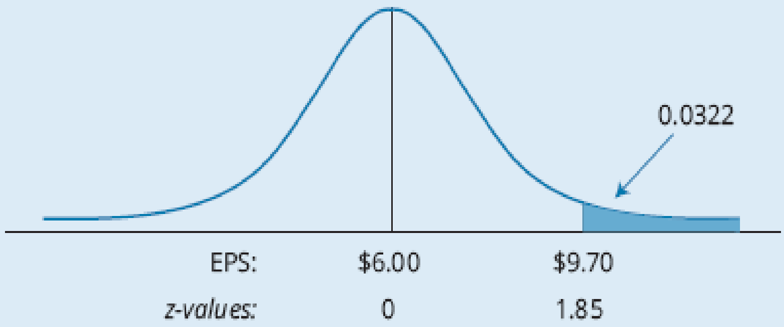
$$z = \frac{(x - \mu)}{\sigma} = \frac{(9.70 - 6)}{2} = 1.85$$

That is, $9.70 is 1.85 standard deviations above the mean EPS value of $6.

From the *z*-table, we have F(1.85) = 0.9678, but this is P(EPS ≤ 9.70). We want P(EPS > 9.70), which is 1 − P(EPS ≤ 9.70).

P(EPS > 9.70) = 1 − 0.9678 = 0.0322, or 3.2%

P(EPS > $9.70)



EXAMPLE: Using the *z*-table (2)

Using the distribution of EPS with μ = $6 and σ = $2 again, what percent of the observed EPS values are likely to be less than $4.10?

Answer:

As shown graphically in the distribution that follows, we want to know P(EPS < $4.10). This requires a two-step approach like the one taken in the preceding example.

First, the corresponding *z*-value must be determined as follows:

$$z = \frac{(\$4.10 - \$6)}{2} = -0.95$$

So, $4.10 is 0.95 standard deviations below the mean of $6.00.

Now, from the *z*-table for negative values in the back of this book, we find that F(−0.95) = 0.1711, or 17.11%.

Finding a Left-Tail Probability



The *z*-table gives us the probability that the outcome will be more than 0.95 standard deviations below the mean.

The Lognormal Distribution

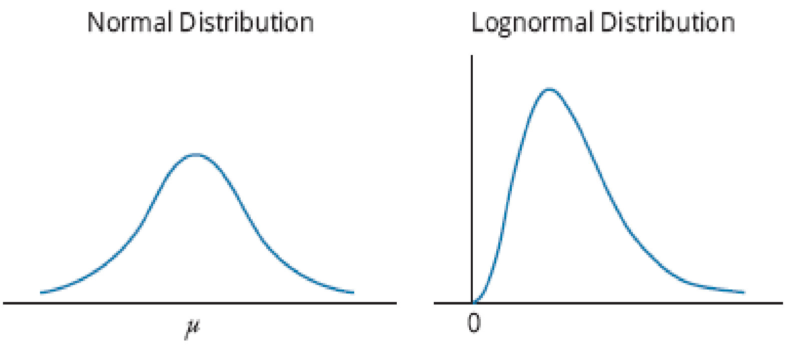
The **lognormal distribution** is generated by the function e^x, where *x* is normally distributed. Because the natural logarithm, ln, of e^x is *x*, the logarithms of lognormally distributed random variables are normally distributed, thus the name.

The PDF for the lognormal distribution is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$$

Figure 14.5 illustrates the differences between a normal distribution and a lognormal distribution.

Figure 14.5: Normal vs. Lognormal Distributions



In Figure 14.5, we can see the following:

- The lognormal distribution is skewed to the right.
- The lognormal distribution is bounded from below by zero so that it is useful for modeling asset prices that never take negative values.

If we used a normal distribution of returns to model asset prices over time, we would admit the possibility of returns less than –100%, which would admit the possibility of asset prices less than zero. Using a lognormal distribution to model *price relatives* avoids this problem. A price relative is just the end-of-period price of the asset divided by the beginning price (S_1/S_0) and is equal to (1 + the holding period return). To get the end-of-period asset price, we can simply multiply the price relative by the beginning-of-period asset price. Because a lognormal distribution takes a minimum value of zero, end-of-period asset prices cannot be less than zero. A price relative of zero corresponds to a holding period return of –100% (i.e., the asset price has gone to zero).

MODULE QUIZ 14.2

1. The probability that a normal random variable will be more than two standard deviations above its mean is:
   A. 0.0217.
   B. 0.0228.
   C. 0.4772.
   D. 0.9772.

2. Which of the following random variables is least likely to be modeled appropriately by a lognormal distribution?
   A. The size of silver particles in a photographic solution.
   B. The number of hours a housefly will live.
   C. The return on a financial security.
   D. The weight of a meteor entering the earth's atmosphere.

MODULE 14.3: STUDENT'S *T*, CHI-SQUARED, AND *F*-DISTRIBUTIONS

Student's *t*-Distribution

Student's *t*-distribution is similar to a normal distribution, but has fatter tails (i.e., a greater proportion of the outcomes are in the tails of the distribution). It is the appropriate distribution to use when constructing confidence intervals based on *small samples* (n < 30) from a population with *unknown variance* and a normal, or approximately normal, distribution. It may also be appropriate to use the *t*-distribution when the population variance is unknown and the sample size is large enough that the central limit theorem will assure that the sampling distribution is approximately normal.

Student's *t*-distribution has the following properties:

- It is symmetrical.
- It is defined by a single parameter, the degrees of freedom (df), where the degrees of freedom are equal to the number of sample observations minus 1, n − 1, for sample means.
- It has a greater probability in the tails (fatter tails) than the normal distribution.
- As the degrees of freedom (the sample size) gets larger, the shape of the *t*-distribution more closely approaches a standard normal distribution.

The degrees of freedom for tests based on sample means are n − 1 because, given the mean, only n − 1 observations can be unique.

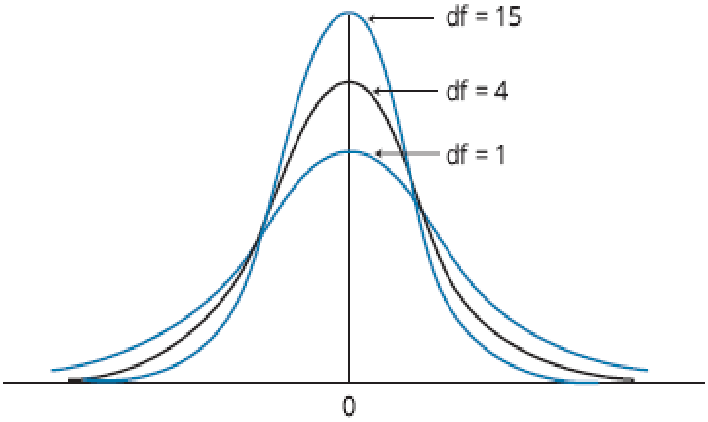
The table in Figure 14.6 contains one-tailed critical values for the *t*-distribution at the 0.05 and 0.025 levels of significance with various degrees of freedom (df). Note that, unlike the *z*-table, the *t*-values are contained within the table and the probabilities are located at the column headings.

Figure 14.6: Table of Critical *t*-Values

	One-Tailed Probabilities, p	
df	p = 0.05	p = 0.025
5	2.015	2.571
10	1.812	2.228
15	1.753	2.131
20	1.725	2.086
40	1.684	2.021
60	1.671	2.000
80	1.664	1.990
100	1.660	1.984
120	1.658	1.980
∞	1.645	1.960

Figure 14.7 illustrates the shapes of the *t*-distribution associated with different degrees of freedom. The tendency is for the *t*-distribution to look more and more like the normal distribution as the degrees of freedom increase. Practically speaking, the greater the degrees of freedom, the greater the percentage of observations near the center of the distribution and the lower the percentage of observations in the tails, which are thinner as degrees of freedom increase. This means that confidence intervals for a random variable that follows a *t*-distribution must be wider than those for a normal distribution, for a given confidence level.

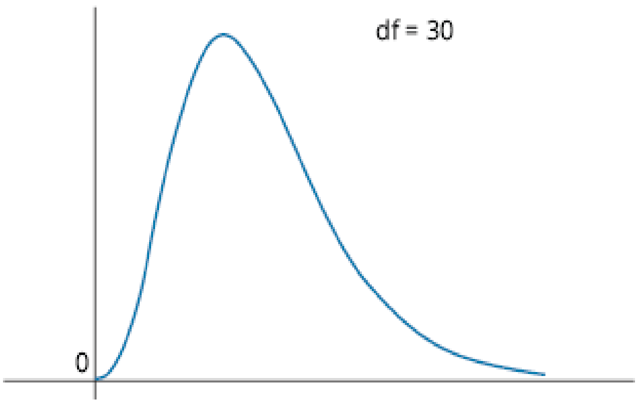
Figure 14.7: *t*-Distributions for Different Degrees of Freedom (df)



The Chi-Squared Distribution

Hypothesis tests concerning population parameters and models of random variables that are always positive are often based on a **chi-squared distribution**, denoted χ^2. The chi-squared distribution is asymmetrical, bounded below by zero, and approaches the normal distribution in shape as the degrees of freedom increase.

Figure 14.8: Chi-Squared Distribution



The chi-squared test statistic, χ^2, with n − 1 degrees of freedom, is computed as:

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2}$$

where:
n = sample size
s^2 = sample variance
σ_0^2 = hypothesized value for the population variance

The chi-squared test compares the test statistic to a critical chi-squared value at a given level of significance to determine whether to reject or fail to reject a null hypothesis.

The *F*-Distribution

Hypotheses concerning the equality of the variances of two populations are tested with an *F*-distributed test statistic. An *F*-distributed test statistic is used when the populations from which samples are drawn are normally distributed and that the samples are independent.

The test statistic for the *F*-test is the ratio of the sample variances. The *F*-statistic is computed as:

$$F = \frac{s_1^2}{s_2^2}$$

where:
s_1^2 = variance of the sample of n_1 observations drawn from Population 1
s_2^2 = variance of the sample of n_2 observations drawn from Population 2

An *F*-distribution is presented in Figure 14.9. As indicated, the *F*-distribution is right-skewed and is truncated at zero on the left-hand side. The shape of the *F*-distribution is determined by *two separate degrees of freedom*, the numerator degrees of freedom, df_1, and the denominator degrees of freedom, df_2.

Note that $n_1 - 1$ and $n_2 - 1$ are the degrees of freedom used to identify the appropriate critical value from the *F*-table (provided in the Appendix).

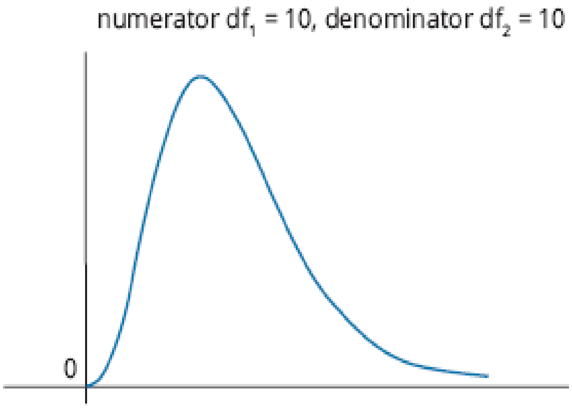
Some additional properties of the *F*-distribution include the following:

- The *F*-distribution approaches the normal distribution as the number of observations increases (just as with the *t*-distribution and chi-squared distribution).
- A random variable's *t*-value squared (t^2) with n − 1 degrees of freedom is *F*-distributed with 1 degree of freedom in the numerator and n − 1 degrees of freedom in the denominator.
- There exists a relationship between the *F*- and chi-squared distributions such that:

  $$F = \frac{\chi^2}{\text{\# of observations in numerator}}$$

  as the # of observations in denominator → ∞

Figure 14.9: *F*-Distribution



The Exponential Distribution

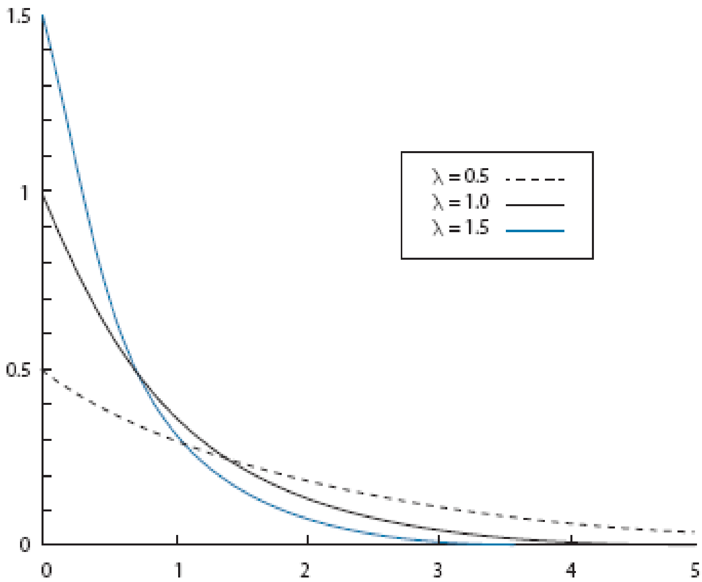
The **exponential distribution** is often used to model waiting times, such as how long it takes an employee to serve a customer or the time it takes a company to default. The PDF for this distribution is as follows:

$$f(x) = \frac{1}{\beta} \times e^{-x/\beta},\ x \geq 0$$

In the previous function, the scale parameter, β, is greater than zero and is the reciprocal of the rate parameter λ (i.e., λ = 1/ β). The rate parameter measures the rate at which it will take an event to occur. In the context of waiting for a company to default, the rate parameter is known as the **hazard rate** and indicates the rate at which default will arrive.

Figure 14.10 displays the PDF of the exponential distribution assuming different values for the rate parameter.

Figure 14.10: Exponential PDF



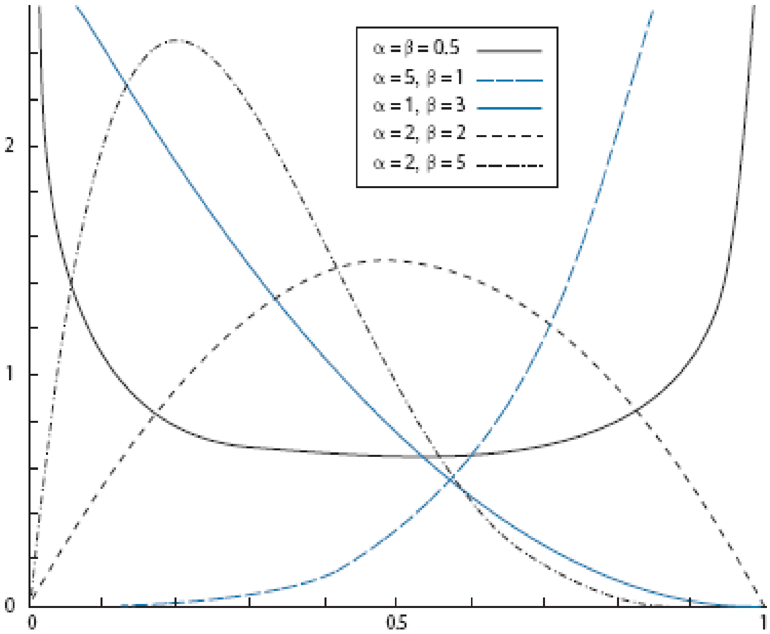
The exponential distribution is able to assess the time it takes a company to default. However, what if we want to evaluate the total number of defaults over a specific period? As it turns out, the number of defaults up to a certain period, N_t, follows a Poisson distribution with a rate parameter equal to t / β.

We can further examine the relationship between the exponential and Poisson distributions by considering the mean and variance of both distributions. Recall that the mean and variance of a Poisson-distributed random variable is equal to λ. As it turns out, the mean of the exponential distribution is equal to 1 / λ, and the variance is equal to $1 / \lambda^2$.

The Beta Distribution

The **beta distribution** can be used for modeling default probabilities and recovery rates. As a result, it is used in some credit risk models such as CreditMetrics®, which will be discussed in the FRM Part II curriculum. The mass of the beta distribution is located between the intervals zero and one. As you can see from Figure 14.11, this distribution can be symmetric or skewed depending on the values of its shape parameters (β and α).

Figure 14.11: Beta PDF



Mixture Distributions

LO 14.b: Describe a mixture distribution and explain the creation and characteristics of mixture distributions.

The distributions discussed in this reading, as well as other distributions, can be combined to create unique PDFs. It may be helpful to create a new distribution if the underlying data you are working with does not currently fit a predetermined distribution. In this case, a newly created distribution may assist with explaining the relevant data.

To illustrate a mixture distribution, suppose that the returns of a stock follow a normal distribution with low volatility 75% of the time and high volatility 25% of the time. Here, we have two normal distributions with the same mean, but different risk levels. To create a mixture distribution from these scenarios, we randomly choose either the

low or high volatility distribution, placing a 75% probability on selecting the low volatility distribution. We then generate a random return from the selected distribution. By repeating this process several times, we will create a probability distribution that reflects both levels of volatility.

Mixture distributions contain elements of both parametric and nonparametric distributions. The distributions used as inputs (i.e., the component distributions) are parametric, while the weights of each distribution within the mixture are nonparametric. The more component distributions used as inputs, the more closely the mixture distribution will follow the actual data. However, more component distributions will make it difficult to draw conclusions given that the newly created distribution will be very specific to the data.

By mixing distributions, it is easy to see how we can alter skewness and kurtosis of the component distributions. Skewness can be changed by combining distributions with different means, and kurtosis can be changed by combining distributions with different variances. Also, by combining distributions that have significantly different means, we can create a mixture distribution with multiple modes (e.g., a bimodal distribution).

Creating a more robust distribution is clearly beneficial to risk managers. Different levels of skew and/or kurtosis can reveal extreme events that were previously difficult to identify. By creating these mixture distributions, we can improve risk models by incorporating the potential for low-frequency, high-severity events.

MODULE QUIZ 14.3

1. The *t*-distribution is the appropriate distribution to use when constructing confidence intervals based on:
   A. large samples from populations with known variance that are nonnormal.
   B. large samples from populations with known variance that are at least approximately normal.
   C. small samples from populations with known variance that are at least approximately normal.
   D. small samples from populations with unknown variance that are at least approximately normal.

2. Which of the following statements about *F*- and chi-squared distributions is least accurate? Both distributions:
   A. are asymmetrical.
   B. are bound by zero on the left.
   C. are defined by degrees of freedom.
   D. have means that are less than their standard deviations.

KEY CONCEPTS

LO 14.a

A continuous uniform distribution is one where the probability of *X* occurring in a possible range is the length of the range relative to the total of all possible values. Letting *a* and *b* be the lower and upper limit of the uniform distribution, respectively, then for $a \le x_1 \le x_2 \le b$,

$$P(x_1 \le X \le x_2) = \frac{(x_2 - x_1)}{(b - a)}$$

The binomial distribution is a discrete probability distribution for a random variable, *X*, that has one of two possible outcomes: success or failure. The probability of a specific number of successes in *n* independent binomial trials is:

$$p(x) = P(X = x) = \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}$$

where p = the probability of success in a given trial

The Poisson random variable, *X*, refers to a specific number of successes per unit. The probability for obtaining *X* successes, given a Poisson distribution with parameter λ, is:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The normal probability distribution has the following characteristics:

- The normal curve is symmetrical and bell-shaped with a single peak at the exact center of the distribution.
- Mean = median = mode, and all are in the exact center of the distribution.
- The normal distribution can be completely defined by its mean and standard deviation because the skew is always 0 and kurtosis is always 3.

A standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. A normal random variable, x, can be normalized (changed to a standard normal, z) with the transformation z = (x – mean of x) / standard deviation of x.

A lognormal distribution exists for random variable *Y*, when $Y = e^X$, and *X* is normally distributed.

The *t*-distribution is similar, but not identical, to the normal distribution in shape—it is defined by the degrees of freedom and has fatter tails. The *t*-distribution is used to construct confidence intervals for the population mean when the population variance is not known.

Degrees of freedom for the *t*-distribution is equal to n − 1; Student's *t*-distribution is closer to the normal distribution when df is greater, and confidence intervals are narrower when df is greater.

The chi-squared distribution is asymmetrical, bounded below by zero, and approaches the normal distribution in shape as the degrees of freedom increase.

The *F*-distribution is right-skewed and is truncated at zero on the left-hand side. The shape of the *F*-distribution is determined by two separate degrees of freedom.

LO 14.b

Mixture distributions combine the concepts of parametric and nonparametric distributions. The component distributions used as inputs are parametric while the

weights of each distribution within the mixture are based on historical data, which is nonparametric.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 14.1

1. **A** The probability of a defective car (*p*) is 0.05; hence, the probability of a nondefective car (q) = 1 – 0.05 = 0.95. Assuming a Poisson distribution:

   $\lambda = np = (3)(0.05) = 0.15$

   Then,

   $$P(X = 1) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(0.15)^1 e^{-0.15}}{1!} = 0.129106$$

   (LO 14.a)

2. **C** Success = having a web page:

   $[6! / 4!(6 - 4)!](0.6)^4(0.4)^{6-4} = 15(0.1296)(0.16) = 0.311$

   (LO 14.a)

3. **B** Since a = 12 and b = 28:

   $$P(15 \le X \le 25) = \frac{(25 - 15)}{(28 - 12)} = \frac{10}{16} = 0.625$$

   (LO 14.a)

Module Quiz 14.2

1. **B** 1 – F(2) = 1 – 0.9772 = 0.0228 (LO 14.a)

2. **C** A lognormally distributed random variable cannot take on values less than zero. The return on a financial security can be negative. The other choices refer to variables that cannot be less than zero. (LO 14.a)

Module Quiz 14.3

1. **D** The *t*-distribution is the appropriate distribution to use when constructing confidence intervals based on small samples from populations with unknown variance that are either normal or approximately normal. (LO 14.a)

2. **D** There is no consistent relationship between the mean and standard deviation of the chi-squared or *F*-distributions. (LO 14.a)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 4.

READING 15

MULTIVARIATE RANDOM VARIABLES

Study Session 4

EXAM FOCUS

This reading covers the dependency of multivariate random variables. For the exam, be prepared to explain and calculate the mean and variance for bivariate random variables. The dependency between the components is important, and you should understand the calculation for covariance and correlation. The marginal and conditional distributions are used to transform bivariate distributions and provide additional insights for finance and risk management. Be able to use these distributions to compute a conditional expectation and conditional moments that summarize the conditional distribution of a random variable.

MODULE 15.1: MARGINAL AND CONDITIONAL DISTRIBUTIONS FOR BIVARIATE DISTRIBUTIONS

Probability Matrices

LO 15.a: Explain how a probability matrix can be used to express a probability mass function.

A **random variable** is an uncertain quantity or number. **Multivariate random variables** are vectors of random variables where a vector is a dimension of *n* random variables. Thus, the study of multivariate random variables includes measurements of dependency between two or more random variables. In this reading, we will examine bivariate random variables or components, which is a special case of the *n* dimension multinomial distribution. The bivariate random variable X is a vector with two components: X_1 and X_2.

A **probability mass function (PMF)** for a bivariate random variable describes the probability that two random variables each take a specific value. The PMF of a bivariate random variable is:

$$f_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

A **probability matrix** illustrates the following properties of a PMF:

- The probability matrix describes the outcome probabilities as a function of the coordinates x_1 and x_2.
- All probabilities are positive or zero and are less than or equal to 1.
- The sum across all possible outcomes for X_1 and X_2 equals 1.

A probability matrix is used to describe the relationship between discrete distributions defined over a finite set of values. The most common application of a discrete bivariate random variable is the trinomial distribution. In this type of example, there are *n* independent trials, and each trial has one of three discrete possible outcomes. The trinomial distribution has three parameters: the number of trials (*n*), the probability of observing outcome 1 (p_1), and the probability of observing outcome 2 (p_2). The sum of all probabilities of each outcome occurring must always equal 100%. Therefore, the probability of the third outcome occurring is found by subtracting p_1 and p_2 from 1 as follows:

$$p_3 = 1 - p_1 - p_2$$

A probability matrix can be created that summarizes the probability of each outcome occurring.

EXAMPLE: Applying a probability matrix

Suppose that a company's common stock return is related to earnings announcements. Earnings announcements are either positive, neutral, or negative and are labeled as 1, 0, and −1, respectively. Assume that the company's monthly stock return must be one of three possible outcomes, −3%, 0%, or 3%. An analyst estimates the probability matrix in Figure 15.1 for earnings announcements and stock returns. **Compute** the probability of a negative earnings announcement.

Figure 15.1: Probability Matrix for Bivariate Random Variables

Earnings (X_2)		Stock Return (X_1) −3%	0%	3%
Negative	−1	25%	15%	0%
Neutral	0	5%	10%	15%
Positive	1	0%	5%	25%

Answer:

The sum of all probabilities in the first row of the probability matrix states that there is a 40% probability of a negative announcement. Also, there is a 25% probability of a negative announcement and a −3% return, a 15% probability of a negative announcement and a 0% return, and a 0% probability of a negative announcement and a 3% return.

Marginal and Conditional Distributions

LO 15.b: Compute the marginal and conditional distributions of a discrete bivariate random variable.

A **marginal distribution** defines the distribution of a *single* component of a bivariate random variable (i.e., a univariate random variable). Thus, the notation for the marginal PMF is the same notation for a univariate random variable:

$$f_{x_1(x_1)} = \sum_{x_2 \in R(X_2)} f_{x_1,x_2}(x_1, x_2)$$

The computation of a marginal distribution can be shown using the previous example of earnings announcements and monthly stock returns. Summing across columns constructs the marginal distribution of the row variables in a probability matrix. Summing across rows constructs the marginal distribution for the column variables in a probability matrix.

EXAMPLE: Marginal distributions

Using the probability matrix in Figure 15.1, **compute** the marginal PMF for the 3% monthly stock return as well as the marginal PMF for a 0% and −3% monthly stock return.

Answer:

The marginal PMF for a 3% monthly stock return, X_1, is calculated by summing the probabilities of all outcomes of 3% across all values based on the earnings announcements, X_2.

$$f_{x_1}(3\%) = \sum_{x_2=(-1,0,1)} f(3\%, x_2) = 0\% + 15\% + 25\% = 40\%$$

The marginal PMF for a 0% and −3% monthly stock return are as follows:

$$f_{x_1}(0\%) = \sum_{x_2=(-1,0,1)} f(0\%, x\ 2) = 15\% + 10\% + 5\% = 30\%$$

$$f_{x_1}(-3\%) = \sum_{x_2=(-1,0,1)} f(-3\%, x\ 2) = 25\% + 5\% + 0\% = 30\%$$

Thus, the complete marginal PMF for monthly stock returns, X_1, is the following.

Return	−3%	0%	3%
Probability	30%	30%	40%

Figure 15.2 illustrates that the sum of the columns and rows in Figure 15.1 are labeled as the marginal PMF for X_1 and X_2. Note that the sum of all possible outcomes for the monthly stock return, X_1, equals 1 at the bottom of Figure 15.2 (30% + 30% + 40% = 100%). Similarly, the sum of all possible outcomes for the earnings announcements at the right of Figure 15.2 must also equal 1 (40% + 30% + 30% = 100%).

Figure 15.2: Marginal PMFs of X_1 and X_2

Earnings (X_2)		Stock Return (X_1)			
		–3%	0%	3%	$fx_2(x_2)$
Negative	–1	25%	15%	0%	40%
Neutral	0	5%	10%	15%	30%
Positive	1	0%	5%	25%	30%
	$fx_1(x_1)$	30%	30%	40%	

A **conditional distribution** sums the probabilities of the outcomes for each component conditional on the other component being a specific value. A conditional PMF is defined based on the conditional probability for a bivariate random variable X_1 given X_2 as:

$$f_{x_1|x_2}(x_1 \mid X_2 = x_2) = \frac{f_{x_1,x_2}(x_1, x_2)}{f_{x_2}(x_2)}$$

The numerator in this equation is the joint probability of two events occurring, and the denominator is the marginal probability that $X_2 = x_2$. Continuing with the previous example, we can determine the three possible outcomes of monthly stock returns given a negative earnings announcement ($x_2 = -1$). When there is a negative earnings announcement, the three probabilities in the first row of the bivariate probability matrix illustrated in Figure 15.2 are 25%, 15%, and 0% for monthly returns of −3%, 0%, and 3%, respectively. These joint probabilities are then divided by the marginal probability of a negative earnings announcement. This is summarized in the upper right-hand corner of Figure 15.2 as 40%.

Thus, the conditional PMF for $X_2 = -1$ is summarized as follows:

Return	–3%	0%	3%
Probability	25% / 40% = 62.5%	15% / 40% = 37.5%	0% / 40% = 0.0%

MODULE QUIZ 15.1

Use the following information to answer Questions 1 and 2.

Suppose a hedge fund manager expects a stock to have three possible returns (–6%, 0%, 6%) following negative, neutral, or positive changes in analyst ratings, respectively. The fund manager constructs the following bivariate probability matrix for the stock.

Ratings (X_2)		Stock Return (X_1)		
		-6%	0%	6%
Negative	-1	30%	15%	0%
Neutral	0	10%	10%	5%
Positive	1	0%	5%	25%

1. What is the marginal probability that the stock has a positive analyst rating?
   A. 10%.

B. 15%.
C. 25%.
D. 30%.

2. What are the conditional probabilities of the three monthly stock returns given that the analyst rating is positive?

	–6%	0%	6%
A.	0.0%	16.7%	83.3%
B.	66.7%	33.3%	0.0%
C.	40.0%	40.0%	0.0%
D.	15.0%	10.0%	5.0%

MODULE 15.2: MOMENTS OF BIVARIATE RANDOM DISTRIBUTIONS

Expectation of a Bivariate Random Function

LO 15.c: Explain how the expectation of a function is computed for a bivariate discrete random variable.

The first moment of a bivariate discrete random variable is referred to as an *expectation* of a function. The expectation of a bivariate random function $g(X_1,X_2)$ is a probability-weighted average of the function of the outcomes $g(x_1,x_2)$ and is expressed as follows:

$$\sum_{x_1 \in R(X_1)}^{n} \sum_{x_2 \in R(X_2)}^{n} g(x_1, x_2) f_{x_1,x_2}(x_1, x_2)$$

The function $g(x_1,x_2)$ depends on x_1 and x_2 but may only be a function of one of the components.

EXAMPLE: Computing expectation of a bivariate random function

Compute the expectation of the function $g(x_1, x_2) = x_1^{x_2}$ using the joint PMF presented in Figure 15.3.

Figure 15.3: Joint PMF for x_1 and x_2

		x_1 = 2	x_1 = 4
x_2	3	25%	5%
	5	45%	25%

Answer:

The expectation is computed as follows:

$$E[g(x_1, x_2)] = \sum_{x_1 \in R(X_1)}^{n} \sum_{x_2 \in R(X_2)}^{n} g(x_1, x_2) f_{x_1,x_2}(x_1, x_2)$$

$$= 2^3(0.25) + 2^5(0.45) + 4^3(0.05) + 4^5(0.25)$$

$$= 2.0 + 14.4 + 3.2 + 256.0 = 275.6$$

Covariance and Correlation Between Random Variables

LO 15.d: Define covariance and explain what it measures.

Expectations of bivariate random variables are used to describe relationships in the same way that they are used to define moments for univariate random variables. For example, the expected return for a stock is used to define the variance of the stock in a univariate random number. The first moment of $X = [X_1, X_2]$ is the expected mean of the components, E[X]. The second moment of a bivariate random *X* has two components and is calculated as a covariance.

Covariance is the expected value of the product of the deviations of the two random variables from their respective expected values. Common notations for the covariance between random variables *X* and *Y* are Cov(X,Y) and σ_{XY}. Covariance measures how two variables move with each other or the dependency between the two variables. The covariance of a multivariate random variable *X* is a 2-by-2 matrix, where the values along one diagonal are the variances of X_1 and X_2. The values along the other diagonal are the covariance between X_1 and X_2. For bivariate random variables, there are two variances and one covariance.

The calculations for variances for bivariate random numbers are analogous to the calculation of dispersion for univariate numbers, where the distance of observations from the expected mean is squared as follows:

$$Var[X_1] = E[(X_1 - E[X_1])^2]$$

The covariance between X_1 and X_2 is calculated as:

$$Cov[X_1,X_2] = E[(X_1 - E[X_1])(X_2 - E[X_2])]$$

$$Cov[X_1,X_2] = E[X_1X_2] - E[X_1]E[X_2]$$

The variances and covariance of two components of *X* are expressed in a 2-by-2 matrix of *X* as:

$$Cov[X] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

To aid in the definition of covariance, consider the returns of a stock and of a put option on the stock. These two returns will have a negative covariance because they move in

opposite directions. The returns of two automotive stocks would likely have a positive covariance, and the returns of a stock and a riskless asset would have a zero covariance because the riskless asset's returns never move, regardless of movements in the stock's return.

LO 15.e: Explain the relationship between the covariance and correlation of two random variables and how these are related to the independence of the two variables.

In practice, covariance is difficult to interpret because it depends on the scales of X_1 and X_2. Thus, it can take on extremely large values, ranging from negative to positive infinity, and, like variance, these values are expressed in terms of squared units.

To make the covariance of two random variables easier to interpret, it may be divided by the product of the bivariate random variables' standard deviations. The resulting value is called the correlation coefficient, or simply, **correlation**. The relationship between covariances, standard deviations, and correlations can be seen in the following expression for the correlation of two bivariate random variables X_1 and X_2:

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var}[X_1]}\sqrt{\text{Var}[X_2]}} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

Correlation measures the strength of the linear relationship between two variables and ranges from −1 to +1 for two variables (i.e., $-1 \leq \text{Corr}(X_1, X_2) \leq +1$). Two variables that are perfectly positively correlated have a correlation coefficient of 1, two variables that are perfectly negatively coefficient have a correlation coefficient of −1, and two variables that are independent have a correlation of 0 (i.e., no linear relationship). However, a correlation of 0 does not necessarily imply independence.

MODULE QUIZ 15.2

1. What is the expectation of the function $g(x_1, x_2) = x_1^{x_2}$ using the following joint PMF?

		x_1	
		3	6
x_2	2	35%	15%
	4	20%	30%

A. 226.4.
B. 358.9.
C. 394.7.
D. 413.6.

2. A hedge fund manager computed the covariances between two bivariate random variables. However, she is having difficulty interpreting the implications of the dependency between the two variables as the scale of the two variables are very different. Which of the following statements will most likely benefit the fund manager when interpreting the dependency for these two bivariate random variables?

A. Compute the correlation by multiplying the covariance of the two variables by the product of the two variables' standard deviations.
B. Disregard the covariance for bivariate random variables as this data is not relevant due to the nature of bivariate random variables.
C. Compute the correlation by dividing the covariance of the two variables by the product of the two variables' standard deviations.
D. Divide the larger scale variables by a common denominator and rerun the estimations of covariance by subtracting each variable's expected mean.

MODULE 15.3: BEHAVIOR OF MOMENTS FOR BIVARIATE RANDOM VARIABLES

Linear Transformations

LO 15.f: Explain the effects of applying linear transformations on the covariance and correlation between two random variables.

There are four important effects of a linear transformation on the covariance of bivariate random variables. The following example illustrates the effects.

Suppose there is a linear relationship between X_1 and X_2 where:

$X_2 = a + bX_1$

The *first effect* of a linear transformation on the covariance of two random variables is that the sign of b determines the correlation between the components. The correlation between X_1 and X_2 will be equal to either

- 1 if $b > 0$,
- 0 if $b = 0$, or
- −1 if $b < 0$.

A *second effect* of linear transformations on covariance is that the amount or scale of a has no effect on the variance, and the scale of b determines the scale or changes in the variance by b^2. This is true because the variance of the linear relationship is equal to:

$b^2Var[X_1]$

A *third effect* of linear transformations on covariance is that the scale of covariance is determined by two variables, *b and d*, as follows:

$Cov[a + bX_1, c + dX_2] = bdCov[X_1,X_2]$

Recall that covariance is defined as the deviation from the expected mean of one random variable multiplied by the deviation from the expected mean of the other random variable. Therefore, location shifts have no impact on the variance or covariance calculations, because only the deviations from the respective means are relevant. However, the scale of each component (b and d) contributes multiplicatively to the change in covariance. We can also extend the first effect and show that the correlation is scale free and is either +1 or −1 when a or b are not equal to zero.

The *fourth effect* of linear transformations on covariance between random variables relates to coskewness and cokurtosis.

Coskewness and **cokurtosis** are cross variable versions of skewness and kurtosis. Interpreting the meaning of coskewness and cokurtosis is not as clear as covariance. However, both coskewness and cokurtosis measure the direction of how one random variable raised to the first power is impacted when the other variable is raised to the second power. For example, stock returns for one variable and volatility of the returns for another variable tend to have negative coskewness. In this case, negative coskewness implies that one variable has a negative return when the other variable has high volatility.

PROFESSOR'S NOTE

The concepts of coskewness and cokurtosis will be illustrated in the next reading (Reading 16).

Variance of Weighted Sum of Bivariate Random Variables

LO 15.g: Compute the variance of a weighted sum of two random variables.

When measuring the variance of two random variables, the covariance or comovement between the two variables is a key component. The variance of two random variables, X_1 and X_2, is computed by summing the individual variances and two times the covariance:

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2] + 2Cov[X_1,X_2]$$

If *a* and *b* represent the weight of investment in asset X_1 and X_2, respectively, then the variance of a two-asset portfolio is computed as follows:

$$Var[aX_1 + bX_2] = a^2Var[X_1] + b^2Var[X_2] + 2abCov[X_1,X_2]$$

In a two-asset portfolio context, this equation is most commonly written as:

$$\sigma_P^2 = w_1^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w_1(1-w)\sigma_{12}$$

The minimum variance portfolio (i.e., optimal risk weight) can then be found as:

$$w^* = \frac{\sigma_{22} - \sigma_{12}}{\sigma_{11} - 2\sigma_{12} + \sigma_{22}}$$

EXAMPLE: Computing variance of a two-asset portfolio

Suppose two assets have a correlation of 0.30. Using the following covariance matrix, **compute** the variance of a two-asset portfolio with 30% in asset 1 and 70% in asset 2.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 18\%^2 & \rho \times 18\% \times 9\% \\ \rho \times 18\% \times 9\% & 9\%^2 \end{pmatrix}$$

Answer:

The variance of this two-asset portfolio is computed as:

$$\begin{aligned} \sigma_P^2 &= (0.30)^2(0.18)^2 + (0.70)^2(0.09)^2 + 2(0.30)(0.70)(0.30 \times 0.18 \times 0.09) \\ &= (0.09)(0.0324) + (0.49)(0.0081) + 0.00204 \\ &= 0.00292 + 0.00397 + 0.00204 = 0.00893 \text{ or } 0.893\% \end{aligned}$$

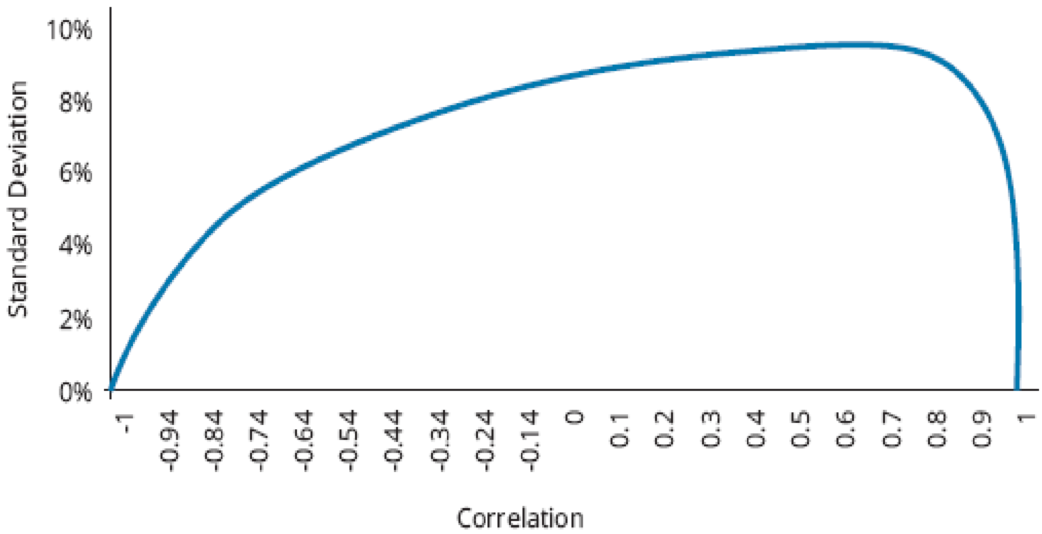
The standard deviation for this two-asset portfolio is 9.45%, which is found by taking the square root of the variance.

Note that the optimal weight of asset 1 with a correlation of 0.30 is approximately 10.5%. The standard deviation of this minimum risk portfolio is approximately 8.8%.

Figure 15.4 illustrates the impact of correlation on the standard deviation for a two-asset portfolio using the optimal (minimum risk) portfolio weight at different correlations between −1 and +1. The optimal weight of asset 1 with a correlation of 0.30 is approximately 10.5%. The standard deviation of this minimum risk portfolio is approximately 8.8%.

We can note a couple of observations from the graph in Figure 15.4. The standard deviation is smallest with strong negative correlations. Second, the graph is asymmetrical because the larger positive correlations result in higher standard deviations (right-hand side of graph) than smaller negative correlations (left-hand side of graph). The reason for the larger correlations is because the optimal weight for the minimum risk portfolio is negative for the largest correlations. This results in the second asset having a weight greater than 1. Unfortunately, with high correlations, the benefits of diversification are limited with more exposure in one asset.

Figure 15.4: Standard Deviation of Two-Asset Portfolio



Conditional Expectations

LO 15.h: Compute the conditional expectation of a component of a bivariate random variable.

In the context of portfolio risk management, a conditional expectation of a random variable is computed based on a specific event occurring. A conditional PMF is used to determine the conditional expectation based on weighted averages. A conditional distribution is defined based on the conditional probability for a bivariate random variable X_1 given X_2.

Suppose a portfolio manager creates a conditional PMF based on earnings announcements, X_2. Earnings announcements can take on three possible outcomes: a positive earnings surprise, $X_2 = 1$; a negative earnings surprise, $X_2 = -1$; or a neutral earnings announcement, $X_2 = 0$. We will return to our previous example in Figure 15.2 and use the same example referred here in Figure 15.5.

Figure 15.5: PMF of Stock Returns, X_1, Given Earnings Announcement, X_2

			Stock Return (X_1)			
			–3%	0%	3%	$f_{X_2}(x_2)$
Earnings (X_2)	Negative	–1	25%	15%	0%	40%
	Neutral	0	5%	10%	15%	30%
	Positive	1	0%	5%	25%	30%
		$f_{X_1}(x_1)$	30%	30%	40%	

The conditional distribution for $f_{X_1|X_2}(x_1 | X_2) = -1$ is summarized as follows:

Return	–3%	0%	3%
Probability	62.5%	37.5%	0.0%

The conditional expectation of the return given that the earnings announcement is negative is then computed as:

$$E[X_1 | X_2 = -1] = -3\% \times 62.5\% + 0\% \times 37.5\% + 3\% \times 0\% = -1.875\%$$

MODULE QUIZ 15.3

1. What is the variance of a two-asset portfolio given the following covariance matrix and a correlation between the two assets of 0.25? Assume the weights in asset 1 and asset 2 are 40% and 60%, respectively.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 10\%^2 & \rho \times 10\% \times 4\% \\ \rho \times 10\% \times 4\% & 4\%^2 \end{pmatrix}$$

A. 0.27%.
B. 0.79%.
C. 1.47%.
D. 2.63%.

2. Suppose a portfolio manager creates a conditional PMF based on analyst ratings, X_2. Analysts' ratings can take on three possible outcomes: an upgrade, $X_2 = 1$; a downgrade, $X_2 = -1$; or a neutral no change rating, $X_2 = 0$. What is the conditional expectation of a return given an analyst upgrade and the following conditional distribution for $f_{X_1|X_2}(x_1 | X_2 = 1)$?

Return	–4%	0%	4%
Probability	12.5%	23.5%	64.0%

A. 2.06%.
B. 3.05%.
C. 4.40%.
D. 11.72%.

MODULE 15.4: INDEPENDENT AND IDENTICALLY DISTRIBUTED RANDOM VARIABLES

LO 15.i: Describe the features of an independent and identically distributed (iid) sequence of random variables.

LO 15.j: Explain how the iid property is helpful in computing the mean and variance of a sum of iid random variables.

Independent and identically distributed (i.i.d.) random variables are generated from a single univariate distribution such as the normal distribution. Features of i.i.d. sequence of random variables include the following:

- Variables are independent of all other components.
- Variables are all from a single univariate distribution.
- Variables all have the same moments.
- Expected value of the sum of n i.i.d. random variables is equal to $n\mu$.
- Variance of the sum of n i.i.d. random variables is equal to $n\sigma^2$.
- Variance of the sum of i.i.d. random variables grows linearly.
- Variance of the average of multiple i.i.d. random variables decreases as n increases.

Determining the mean and variance of i.i.d. random variables is relatively easy because the variables are independent and have the same moments. The expected value of the sum of n i.i.d. random variables is simply equal to $n\mu$. All i.i.d. random variables are from the same univariate distribution and thus have the same mean, μ. The expectation of a sum is always the sum of the expectations. In this case, we assume all variables are identical. Thus, the sum of the means is simply a linear scale based on n as follows:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n\mu$$

Similarly, the variance of n i.i.d. random variables is equal to $n\sigma^2$. This result is only true if the variables are independent of each other in addition to identical. This can be illustrated with the following equations.

The variance of i.i.d. random variables is computed as:

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i] + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} \text{Cov}(X_j, X_k)$$

Because all variables are independent, the covariances of all variables must be equal to zero.

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \sigma^2 + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} 0$$

This results in the second term of the previous equation dropping out, and we are simply left with the sum of all variances. Since all i.i.d. random variables have the same expected mean and variance, the variance of a sum of i.i.d. random variables is equal to $n\sigma^2$.

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2$$

The variance of the sum of multiple random variables grows linearly based on n. Thus, for two i.i.d. random variables, X_1 and X_2, the variance will be $2\sigma^2$.

$$\text{Var}[X_1 + X_2] = 2\sigma^2$$

An important implication for this when estimating unknown parameters is that the variance of the average reduces as n increases. In other words, with a larger n, the expected average will be closer to the true unknown mean, μ.

MODULE QUIZ 15.4

1. Which of the following statements regarding the sums of i.i.d. normal random variables is incorrect?
   A. The sums of i.i.d. normal random variables are normally distributed.
   B. The expected value of a sum of three i.i.d. random variables is equal to 3μ.
   C. The variance of the sum of four i.i.d. random variables is equal to $6\sigma^2$.
   D. The variance of the sum of i.i.d. random variables grows linearly.

2. The variance of the average of multiple i.i.d. random variables:
   A. increases as n increases.
   B. decreases as n increases.
   C. increases if the covariance is negative as n increases.
   D. decreases if the covariance is negative as n increases.

KEY CONCEPTS

LO 15.a

A probability matrix of a discrete bivariate random variable distribution describes the outcome probabilities as a function of the coordinates x_1 and x_2. All probabilities in the matrix are positive or zero, are less than or equal to 1, and the sum across all possible outcomes for X_1 and X_2 equals 1.

LO 15.b

A marginal distribution defines the distribution of a single component of a bivariate random variable (i.e., a univariate random variable). A conditional distribution sums the probabilities of the outcomes for each component conditional on the other component being a specific value.

LO 15.c

The expectation of a bivariate random function $g(X_1, X_2)$ is a probability-weighted average of the function of the outcomes $g(x_1, x_2)$.

LO 15.d

Covariance is the expected value of the product of the deviations of the two random variables from their respective expected values. It measures how two variables move with each other.

LO 15.e

The correlation coefficient is a statistical measure that standardizes the covariance as follows:

$$\text{Corr}(X_1, X_2) = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

LO 15.f

The first effect of a linear transformation on the covariance of two random variables is that *b* determines the correlation between the components. The correlation between X_1 and X_2 will be 1 if $b > 0$, 0 if $b = 0$, and -1 if $b < 0$.

A second effect of linear transformations on covariance is that the amount or scale of *a* has no effect on the variance, and the scale of *b* determines the scale or changes in the variance by b^2.

A third effect of linear transformations on covariance is that the scale of covariance is determined by two variables, *b* and *d*, as follows:

$$\text{Cov}[a + bX_1, c + dX_2] = bd\text{Cov}[X_1, X_2]$$

The fourth effect of linear transformations on covariance between random variables relates to coskewness and cokurtosis.

LO 15.g

The variance of a two-asset portfolio, X_1 and X_2, with weights of *a* and *b*, respectively is:

$$\text{Var}[aX_1 + bX_2] = a^2\text{Var}[X_1] + b^2\text{Var}[X_2] + 2ab\text{Cov}[X_1, X_2]$$

LO 15.h

In the context of portfolio risk management, a conditional expectation of a random variable is computed based on a specific event occurring. A conditional distribution is defined based on the conditional probability for a bivariate random variable X_1 given X_2.

LO 15.i

Independent and identically distributed (i.i.d.) random variables

- are independent of all other components,
- are all from a single univariate distribution, and
- all have the same moments.

LO 15.j

Features of the sum of *n* i.i.d. random variables include the following:

- The expected value of the sum of *n* i.i.d. random variables is equal to $n\mu$.
- The variance of the sum of *n* i.i.d. random variables is equal to $n\sigma^2$.
- The variance of the sum of i.i.d. random variables grows linearly.
- The variance of the average of multiple i.i.d. random variables decreases as *n* increases.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 15.1

1. **D** The marginal distribution for a positive analyst rating is computed by summing the third row consisting of all possible outcomes of a positive rating as follows:

   $f_{X_2}(1) = 0\% + 5\% + 25\% = 30\%$ (LO 15.b)

2. **A** A conditional distribution is defined based on the conditional probability for a bivariate random variable X_1 given X_2. All possible outcomes of a positive analyst rating are found in the third row of the bivariate probability matrix ($x_2 = 1$) as 0%, 5%, and 25% for monthly returns of –6%, 0%, and 6%, respectively. These joint probabilities are then divided by the marginal probability of a positive analyst rating, which is computed as 0% + 5% + 25% = 30%. Thus, the conditional distribution for $X_2 = 1$ is computed as 0% / 30%, 5% / 30%, and 25% / 30% and summarized as follows:

Return	–6%	0%	6%
Probability	0%	16.7%	83.3%

   (LO 15.b)

Module Quiz 15.2

1. **D** The expectation is computed as follows:

   $$E[g(x_1, x_2)] = 3^2(0.35) + 3^4(0.20) + 6^2(0.15) + 6^4(0.30)$$

   = 3.15 + 16.20 + 5.40 + 388.80 = 413.55 (LO 15.c)

2. **C** Correlation will standardize the data and remove the difficulty in interpreting the scale difference between variables. Correlation is determined by dividing the covariance of the two variables by the product to the two variables' standard deviations. The formula for correlation is as follows:

   $$\text{Corr}(X_1, X_2) = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

(LO 15.e)

Module Quiz 15.3

1. **A** The variance of this two-asset portfolio is computed as:

$$\sigma_{12}^2 = (0.40)^2(0.10)^2 + (0.60)^2(0.04)^2 + 2(0.40)(0.60)(0.25 \times 0.10 \times 0.04)$$
$$= (0.16)(0.01) + (0.36)(0.0016) + 0.00048$$
$$= 0.00160 + 0.00058 + 0.00048 = 0.00266 \text{ or } 0.27\%$$

(LO 15.g)

2. **A** The conditional expectation of the return given a positive analyst upgrade is computed as:

$$E[X_1 | X_2 = 1] = -4\% \times 12.5\% + 0\% \times 23.5\% + 4\% \times 64.0\%$$
$$= -0.005 + 0.0 + 0.0256 = 0.0206 \text{ or } 2.06\%$$

(LO 15.h)

Module Quiz 15.4

1. **C** The variance of the sum of *n* i.i.d. random variables is equal to $n\sigma^2$. Thus, for four i.i.d. random variables, the sum of the variance would be equal to $4\sigma^2$. The covariance terms are all equal to zero because all variables are independent. (LO 15.i)

2. **B** The variance of the average of multiple i.i.d. random variables *decreases* as *n* increases. The covariance of i.i.d. random variables is always zero. (LO 15.j)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 5.

READING 16

SAMPLE MOMENTS

Study Session 5

EXAM FOCUS

This reading explains how sample moments (mean, variance, skewness, and kurtosis) are used to estimate the true population moments for data generated from independent and identically distributed (i.i.d.) random variables. For the exam, be able to estimate these sample moments and explain the differences from population moments. Also, be prepared to discuss what makes estimators biased, unbiased, and consistent. In addition, be able to discuss the law of large numbers (LLN) and the central limit theorem (CLT). Lastly, be prepared to contrast the advantages of estimating quantiles to traditional measures of dispersion.

MODULE 16.1: ESTIMATING MEAN, VARIANCE, AND STANDARD DEVIATION

LO 16.a: Estimate the mean, variance, and standard deviation using sample data.

The **sample mean**, $\hat{\mu}$, is estimated by dividing the sum of all the values in a sample of a population, ΣX, by the number of observations in the sample, n. It is used to make *inferences* about the population mean. The sample mean is expressed as:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The sum of the deviations of each observation in the data set from the mean is always zero.

The arithmetic mean is the only measure of central tendency for which the sum of the deviations from the mean is zero. Mathematically, this property can be expressed as follows:

$$\text{sum of mean deviations} = \sum_{i=1}^{n}\left(X_i - \hat{\mu}\right) = 0$$

The deviations are squared to estimate the **variance** of the sample. The *biased sample estimator* of variance for a sample of n i.i.d. random variables X_i is computed:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2$$

The square root of the variance is called the **standard deviation**. The variance and standard deviation measure the extent of the dispersion in the values of the random variable around the mean.

EXAMPLE: Estimating the mean, variance, and standard deviation with sample data

Assume you are evaluating the stock of Alpha Corporation. You have calculated the stock returns for Alpha Corporation over the last five years to develop the following sample data set. Given this information, **calculate** the sample mean, variance, and standard deviation.

Data set: 24%, 34%, 18%, 54%, 10%

Answer:

$$\hat{\mu} = \text{sample mean} = \frac{0.24 + 0.34 + 0.18 + 0.54 + 0.10}{5} = 0.28 = 28.0\%$$

The calculation of the sample variance can be computed using the following table:

X_i	Mean	Deviation	Squared Deviation
0.24	0.28	−0.04	0.0016
0.34	0.28	0.06	0.0036
0.18	0.28	−0.10	0.01
0.54	0.28	0.26	0.0676
0.10	0.28	−0.18	0.0324
			0.1152

From the table, the first step is to compute the deviation from the mean. In the third column, the mean is subtracted from the observed value, X_i. In the fourth column, the deviations from the mean in the third column are squared. The sum of all squared deviations is equal to 0.1152. This amount is then divided by the number of observations to compute the variance of 0.023 (= 0.1152/5).

The biased standard deviation for this sample is then computed as:

$\sqrt{0.023} = 0.1517$ or 15.17%

The calculations in the previous example result in a biased estimate of the variance and standard deviation. Because the bias is known, the estimate of variance and standard deviation should be divided by (n − 1) and not n. (This will be discussed later in this reading.)

Therefore, the *unbiased* estimate for variance is computed by dividing the sum of all squared deviations by (n − 1).

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

Given the data in the previous example, this results in an unbiased estimate of 0.0288 for the variance (= 0.1152/4). The unbiased estimate of the standard deviation is then 0.1697 or 16.97%.

PROFESSOR'S NOTE

Unless you are specifically instructed on the exam to compute a biased variance, you should always compute the unbiased variance by dividing by (n – 1).

Population and Sample Moments

LO 16.b: Explain the difference between a population moment and a sample moment.

Measures of **central tendency** identify the center, or average, of a data set. This central point can then be used to represent the typical, or expected, value in the data set. The first moment of the distribution of data is the mean.

To compute the **population mean**, μ, all the observed values in the population are summed and divided by the number of observations in the population, *N*. Note that the population mean is unique in that a given population has only one mean. The population mean is expressed as:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

The population mean is unknown because not all of the random numbers of the population are observable. Therefore, we create samples of data to estimate the true population mean. The hat notation above the μ, denotes that the **sample mean**, $\hat{\mu}$, is an estimate of the true mean.

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The sample mean is an estimate based on a known data set where all data points are observable. Thus, the sample mean is simply an estimate of the true population mean. Note the use of *n*, the sample size, versus *N*, the population size.

The population mean and sample mean are both examples of **arithmetic means**. The arithmetic mean is the sum of the observed values divided by the number of observations. It is the most widely used measure of central tendency and has the following properties:

- All interval and ratio data sets have an arithmetic mean.

- All data values are considered and included in the arithmetic mean computation.
- A data set has only one arithmetic mean (i.e., the arithmetic mean is unique).

The following example illustrates the difference between the sample mean and the population mean.

EXAMPLE: Estimating the mean with different sample sizes

Assume you and your research assistant are evaluating the stock of Beta Corporation. You have calculated the stock returns for Beta Corporation over the last 12 years to develop the following data set. Your research assistant has decided to conduct his analysis using only the returns for the five most recent years, which are displayed as the bold numbers in the data set. Given this information, **calculate** the two sample means and discuss the population mean.

Data set: 12%, **25**%, **34**%, 15%, **19**%, 44%, **54**%, 33%, 22%, 28%, **17**%, 24%

Answer:

$\hat{\mu}$ = 1st sample mean =

$$\frac{12 + 25 + 34 + 15 + 19 + 44 + 54 + 33 + 22 + 28 + 17 + 24}{12} = 27.25\%$$

$$\hat{\mu} = 2^{nd} \text{ sample mean } = \frac{25 + 34 + 19 + 54 + 17}{5} = 29.8\%$$

The population mean is the expected value of all possible random returns for the company. Because all possible random observations of returns are not observable, sample estimates are used to estimate the true population mean. A larger sample size results in an estimate that is closer to the true unobservable population mean.

Unusually large or small values can have a disproportionate effect on the computed value for the arithmetic mean. For example, the mean of 1, 2, 3, and 50 is 14 and is not a good indication of what the individual data values really are. On the positive side, the arithmetic mean uses all the information available about the observations. The arithmetic mean of a sample from a population is the best estimate of both the true mean of the sample and the value of the next observation.

Variance and Standard Deviation

The mean and variance of a distribution are defined as the first and second moments of the distribution, respectively. The variance of an estimator or sample mean can be calculated using the standard properties of random variables as the sum of the variances and covariances:

$$\text{Var}[\hat{\mu}] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} \text{Var}(X_i) + \text{Cov}\right]$$

All covariances are 0, because the X_i are all uncorrelated random variables that are i.i.d. This results in the second term in the brackets dropping out. The estimate for variance

then simplifies to:

$$\text{Var}[\hat{\mu}] = \frac{1}{n^2}\left[\sum_{i=1}^{n}\text{Var}(X_i)+\text{Cov}\right] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Thus, the variance of the mean estimator depends on the variance of the sample data and the number of observations. If data is more variable, then it is more difficult to estimate the true variance. The variance of the mean estimator will decrease when the size of the sample or number of observations is increased. Therefore, a larger sample size helps to reduce the difference between the estimated variance and the true variance of the population.

The **variance** of a random variable is defined as:

$$\sigma^2 = \text{Var}[X] = E[(X - E[X])^2]$$

The *population moments* are transformed into *estimator moments* by replacing the expected operator of a random variable, E[X], with an averaging operator that divides by the number of observations, *n*, such that:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2$$

Point Estimates and Estimators

LO 16.c: Distinguish between an estimator and an estimate.

Sample parameters can be used to draw conclusions about true population parameters which are unknown. **Point estimates** are single (sample) values used to estimate population parameters, and the formula used to compute a point estimate is known as an **estimator**. The hat notation, $\hat{\mu}$, in the following formula denotes that the estimator formula of the mean is used to estimate the true unknown mean parameter, μ.

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample data is then used instead of random data from a population, X_i. The mean estimator is a formula that transforms data into an estimate of the true population mean using observed data from a sample of the population.

Biased Estimators

LO 16.d: Describe the bias of an estimator and explain what the bias measures.

The bias of an estimator measures the difference between the expected value of the estimator, $E[\hat{\theta}]$, and the true population value, θ. Therefore, the estimator bias is computed as:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

The expected value of the mean estimator is equal to the true population mean. When X_i consists of i.i.d. random variables, the mean estimator is equal to the true population mean, μ. The following equation illustrates that the mean estimator bias is zero, because the expected mean estimator is equal to the true population mean.

$$\text{Bias}(\hat{\mu}) = E[\hat{\mu}] - \mu = \mu - \mu = 0$$

Therefore, the *sample mean* is an *unbiased estimator*. Conversely, the *sample variance* is a *biased estimator*. The bias for the estimator is based on the sample size *n*. The expected sample variance, $E[\hat{\sigma}^2]$, is a function of the true population variance, σ^2, and the number of observations, *n*:

$$E[\hat{\sigma}^2] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

The sample variance is then computed as:

$$\text{Bias}(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \hat{\sigma}^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{\sigma^2}{n}$$

Thus, when the sample size *n* is large, the bias is small. The fact that the bias is known allows us to determine an unbiased estimator for the sample variance as:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$$

Note that based on the mathematical theory behind statistical procedures, the use of the entire number of sample observations, *n*, instead of n − 1 as the divisor in the computation of s^2, will systematically underestimate the population parameter, σ^2, particularly for small sample sizes. This systematic underestimation causes the sample variance to be a biased estimator of the population variance. Using n − 1 instead of *n* in the denominator, however, improves the statistical properties of s^2 as an estimator of σ^2. Thus, s^2, as expressed in the equation, is considered to be an unbiased estimator of σ^2.

Best Linear Unbiased Estimator

LO 16.e: Explain what is meant by the statement that the mean estimator is BLUE.

The **best linear unbiased estimator (BLUE)** is the best estimator of the population mean available because it has the minimum variance of any linear unbiased estimator. When data is i.i.d., the sample mean is considered to be BLUE.

The following equation denotes how linear estimators of the mean are computed:

$$\hat{\mu} = \sum_{i=1}^{n} w_i X_i$$

Where w_i are the weights that are independent of X_i (i.e., $w_i = 1/n$).

Because the observations are equally likely, the weights are all equal to $1/n$. An *unbiased estimator* is one for which the expected value of the estimator is equal to the parameter you are trying to estimate. For example, the sample mean is an unbiased estimator of the population mean, because the expected value of the sample mean is equal to the population mean.

Note that there may be other nonlinear estimators that are better at estimating the true parameters of a distribution. For example, maximum likelihood estimators of the population mean may be more accurate. However, these estimators are nonlinear and are often biased in finite samples.

MODULE QUIZ 16.1

1. A risk manager gathers the following sample data to analyze annual returns for an asset: 12%, 25%, and –1%. He wants to compute the best unbiased estimator of the true population mean and standard deviation. The manager's estimate of the standard deviation for this asset should be closest to:
   A. 0.0111.
   B. 0.0133.
   C. 0.1054.
   D. 0.1300.
2. The sample mean is an unbiased estimator of the population mean because the:
   A. sampling distribution of the sample mean is normal.
   B. expected value of the sample mean is equal to the population mean.
   C. sample mean provides a more accurate estimate of the population mean as the sample size increases.
   D. sampling distribution of the sample mean has the smallest variance of any other unbiased estimators of the population mean.

MODULE 16.2: ESTIMATING MOMENTS OF THE DISTRIBUTION

LO 16.f: Describe the consistency of an estimator and explain the usefulness of this concept.

LO 16.g: Explain how the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) apply to the sample mean.

Law of Large Numbers

If the **law of large numbers (LLN)** applies to estimators, then the estimators are consistent. The first property of a *consistent estimator* is that as the sample size increases, the finite sample bias is reduced to zero. The second property of a consistent estimator is as the sample size increases, the variance of the estimator approaches zero. The properties of consistency ensure that estimates from large samples have small deviations from the true population mean. This is an important concept that ensures that the estimate of the mean and variance will be very close to the true mean and variance of the population in large sample sizes. Thus, increasing the sample size results in better estimates of the true population distribution.

Central Limit Theorem

The **central limit theorem (CLT)** states that for simple random samples of size *n* from a population with a mean μ and a finite variance σ^2, the sampling distribution of the sample mean, μ, approaches a normal probability distribution with mean μ and variance equal to σ^2/n as the sample size becomes large. The CLT requires only one additional assumption from the LLN that the variance is finite. The LLN only requires the assumption that the mean is finite. In addition, the CLT does not require assumptions about the distribution of the random variables of the population. No assumption regarding the underlying distribution of the population is necessary because, when the sample size is large, the sums of **i.i.d. random variables** (the individual items drawn for the sample) will be normally distributed.

The CLT is extremely useful because the normal distribution is easily applied in testing hypotheses and constructing confidence intervals. Specific inferences about the population mean can be made from the sample mean, regardless of the population's distribution, as long as the sample size is sufficiently large, which usually means $n \geq 30$. As the sample size increases, the sample distribution appears to be more normally distributed.

Important properties of the central limit theorem include the following:

- If the sample size *n* is sufficiently large, the sampling distribution of the sample means will be approximately normal. Remember what's going on here: random samples of size *n* are repeatedly being taken from an overall larger population. Each of these random samples has its own mean, which is itself a random variable, and this set of sample means has a distribution that is approximately normal.
- The mean of the population, μ, and the mean of the distribution of all possible sample means are equal.
- The variance of the distribution of sample means is σ^2/n, the population variance divided by the sample size. Thus, it approaches zero as the sample size increases.

Skewness and Kurtosis

LO 16.h: Estimate and interpret the skewness and kurtosis of a random variable.

The **skewness** statistic is the standardized third central moment of the distribution. Skewness (sometimes called *relative skewness*) refers to the extent to which the distribution of data is not symmetric around its mean. It is calculated as:

$$\text{Skewness}(X) = \frac{E[(X - E[X])^3]}{E[(X - E[X])^2]^{3/2}} = \frac{\mu_3}{\sigma^3}$$

The estimator for the third moment is computed as:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^3}{\hat{\sigma}^3}$$

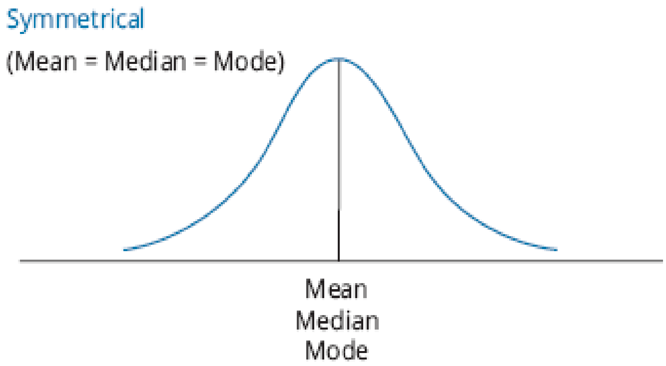
Nonsymmetrical distributions may be either positively or negatively skewed and result from the occurrence of outliers in the data set. **Outliers** are observations with extraordinarily large values, either positive or negative.

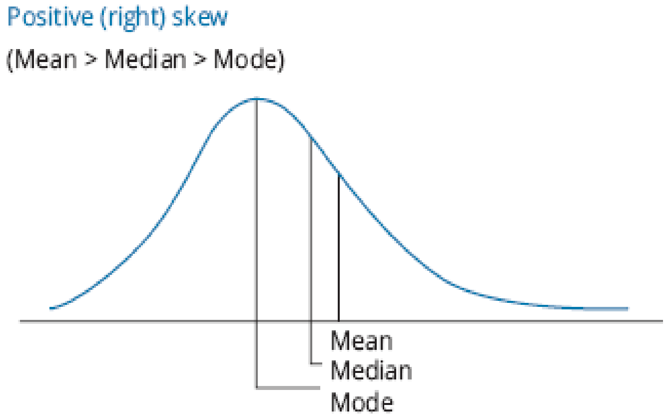
- A *positively skewed* distribution is characterized by many outliers in the upper region, or right tail. A positively skewed distribution is said to be skewed right because of its relatively long upper (right) tail.
- A *negatively skewed* distribution has a disproportionately large amount of outliers that fall within its lower (left) tail. A negatively skewed distribution is said to be skewed left because of its long lower tail.

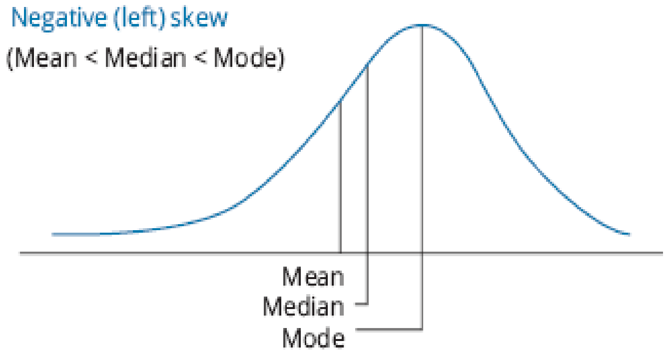
Figure 16.1 illustrates that skewness affects the location of the mean, median, and mode of a distribution. The mean is the arithmetic average, the median is the middle of the ranked data in order, and the mode is the most probable outcome.

- For a symmetrical distribution, the mean, median, and mode are equal.
- For a positively skewed, unimodal distribution, the mode is less than the median, which is less than the mean. The mean is affected by outliers; in a positively skewed distribution, there are large, positive outliers that will tend to pull the mean upward, or more positive. An example of a positively skewed distribution is that of housing prices. Suppose you live in a neighborhood with 100 homes; 99 of them sell for $100,000, and one sells for $1 million. The median and the mode will be $100,000, but the mean will be $109,000. Hence, the mean has been pulled upward (to the right) by the existence of one home (outlier) in the neighborhood.
- For a negatively skewed, unimodal distribution, the mean is less than the median, which is less than the mode. In this case, there are large, negative outliers that tend to pull the mean downward (to the left).

Figure 16.1: Effect of Skewness on Mean, Median, and Mode







The **kurtosis** statistic is the standardized fourth central moment of the distribution. Kurtosis refers to how fat or thin the tails are in the data distribution and is calculated as:

$$\text{Kurtosis}(X) = \frac{E[(X - E[X])^4]}{E[(X - E[X])^2]^2} = \frac{\mu_4}{\sigma^4}$$

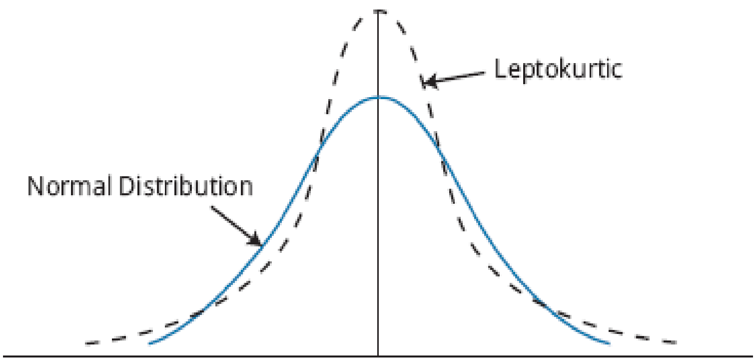
The estimator for the fourth moment is computed as:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^4}{\hat{\sigma}^4}$$

Kurtosis for the normal distribution equals 3. Distributions with a kurtosis greater than 3 are referred to as heavy-tailed or fat-tailed. Many software packages report **excess kurtosis** for any distribution as excess kurtosis = kurtosis − 3. Thus, a normal distribution has excess kurtosis equal to zero.

Figure 16.2 illustrates that relative to a normal distribution, a **leptokurtic distribution** will have a greater percentage of extremely large deviations from the mean (i.e., fat tails). This means there is a relatively greater probability of an observed value being far from the mean. With regard to an investment returns distribution, a greater likelihood of a large deviation from the mean return is often perceived as an increase in risk. Note that a distribution that has thinner tails than a normal distribution if referred to as a **platykurtic distribution**.

Figure 16.2: Kurtosis



Kurtosis is critical in a risk management setting. Most research about the distribution of securities returns has shown that returns are not normally distributed. Actual securities returns tend to exhibit both skewness and kurtosis. Skewness and kurtosis are critical concepts for risk management because when securities returns are modeled using an assumed normal distribution, the predictions from the models will not take into account the potential for extremely large, negative outcomes. In fact, most risk managers put very little emphasis on the mean and standard deviation of a distribution and focus more on the distribution of returns in the tails of the distribution—that is where the risk is. In general, greater positive kurtosis and more negative skew in return distributions indicates increased risk.

Median and Quantile Estimates

LO 16.i: Use sample data to estimate quantiles, including the median.

The **median** is the 50th percentile or midpoint of a data set when the data is arranged in ascending or descending order. It is similar to the mean because both measure the central tendency of the data. If the data is symmetrical then the mean and median are the same when half the observations lie above the median and half are below.

The median is important because the arithmetic mean can be affected by extremely large or small values (outliers). When this occurs, the median is a better measure of central tendency than the mean because it is not affected by extreme values that may possibly be errors in the data.

Estimating quantiles

To determine the median and other quantiles, arrange the data from the highest to the lowest value, or lowest to highest value, and find the middle observation. The middle of

the observations will depend on whether the total sample size is an odd or even number.

The median is estimated when the total number of observations in the sample size is odd as:

$$\text{Median}(x) = x_{(n+1)/2}$$

The median is estimated when the total number of observations in the sample size is even as:

$$\text{Median}(x) = 0.5 \times (x_{n/2} + x_{n/2+1})$$

EXAMPLE: **Estimating the median using an odd number of observations**

What is the median return for five portfolio managers with 10-year annualized total returns of: 30%, 15%, 25%, 21%, and 23%?

Answer:

First, arrange the returns in descending order.

30%, 25%, 23%, 21%, 15%

Then, select the observation that has an equal number of observations above and below it—the one in the middle. For the given data set, the third observation, 23%, is the median value.

EXAMPLE: **Estimating the median using an even number of observations**

Suppose we add a sixth manager to the previous example with a return of 28%. What is the median return?

Answer:

Arranging the returns in descending order gives us:

30%, 28%, 25%, 23%, 21%, 15%

With an even number of observations, there is no single middle value. The median value in this case is the arithmetic mean of the two middle observations, 25% and 23%. Thus, the median return for the six managers is 24.0% = 0.5(25 + 23).

Estimating quartiles

In addition to the median, the two most commonly reported quantiles are the 25th and 75th quantiles. The estimation procedure for these quantiles is similar to the median process. The data is first sorted and then the α-quantile is estimated using the data point in location $\alpha \times n$. If this data value is not an integer value, then the general rule is to average the points immediately above and below $\alpha \times n$.

An **interquartile range (IQR)** is a measure of dispersion from the median similar to the measure of standard deviation from the mean. A common IQR is the range from the 25th to 75th quartile. These measures are useful in determining the symmetry of the distribution and weight of the tails.

There are two properties of quantiles that make them valuable in data analysis:

- The interpretation of the quantiles is easy because they have the same units as the sample data. In other words, there is a 25% probability of obtaining an observation that is in the quartile.
- Quantiles are a robust measure for outliers or extreme values from the mean. In other words, the median and the IQR are not impacted by outliers. Conversely, the mean is impacted by outliers.

Mean of Two Random Variables

LO 16.j: Estimate the mean of two variables and apply the CLT.

The mean of two random variables is estimated the same way as the mean for individual variables. The arithmetic average of the sample is determined by adding up all values and dividing by the number of observations in the sample, *n*. Thus, the formulas for estimating the means of two random variables, X_i and Y_i are:

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

If the data is i.i.d., then the CLT applies to both estimators. If the two mean estimators are considered as a bivariate mean estimate, μ, we can apply the CLT and examine the joint behavior by stacking the two mean estimators into a vector:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{bmatrix}$$

If the multivariate random variable Z = [X, Y] is i.i.d., then the 2 by 1 vector is asymptotically normally distributed (i.e., the estimator converges to the normal distribution as sample size increases).

Covariance and Correlation Between Random Variables

LO 16.k: Estimate the covariance and correlation between two random variables.

The **covariance** between two random variables is a statistical measure of the degree to which the two variables move together. The covariance captures the linear relationship between one variable and another. A positive covariance indicates that the variables tend to move together; a negative covariance indicates that the variables tend to move in opposite directions. Because we will be mostly concerned with the covariance of

asset returns, the following formula has been written in terms of the covariance of the return of asset *X*, and the return of asset *Y*:

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

This equation simplifies to:

$$Cov(X,Y) = E(X,Y) - E(X) \times E(Y)$$

The sample covariance estimator can be calculated as:

$$\text{sample } Cov_{XY} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{n-1}$$

EXAMPLE: Covariance

Assume that the economy can be in three possible states (S) next year: boom, normal, or slow economic growth. An expert source has calculated that P(boom) = 0.30, P(normal) = 0.50, and P(slow) = 0.20. The returns for Stock A, R_A, and Stock B, R_B, under each of the economic states are provided in the following table. What is the covariance of the returns for Stock A and Stock B?

Answer:

First, the expected returns for each of the stocks must be determined.

$$E(R_A) = (0.3)(0.20) + (0.5)(0.12) + (0.2)(0.05) = 0.13$$

$$E(R_B) = (0.3)(0.30) + (0.5)(0.10) + (0.2)(0.00) = 0.14$$

The covariance can now be computed using the procedure described in the following table.

Covariance Computation

Event	P(S)	R_A	R_B	$P(S) \times [R_A - E(R_A)] \times [R_B - E(R_B)]$	
Boom	0.3	0.20	0.30	(0.3)(0.2 – 0.13)(0.3 – 0.14)	= 0.00336
Normal	0.5	0.12	0.10	(0.5)(0.12 – 0.13)(0.1 – 0.14)	= 0.00020
Slow	0.2	0.05	0.00	(0.2)(0.05 – 0.13)(0 – 0.14)	= 0.00224
	$Cov(R_A, R_B) = \Sigma P(S) \times [R_A - E(R_A)] \times [R_B - E(R_B)] = 0.00580$				

The actual value of the covariance is not very meaningful because its measurement is extremely sensitive to the scale of the two variables. Also, the covariance may range from negative to positive infinity and it is presented in terms of squared units (e.g., percent squared). For these reasons, we take the additional step of calculating the **correlation** coefficient, which converts the covariance into a measure that is easier to interpret:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

EXAMPLE: Correlation

Using our previous example, **compute** and **interpret** the correlation of the returns for Stocks A and B, given that $\sigma^2(R_A) = 0.0028$ and $\sigma^2(R_B) = 0.0124$ and recalling that $Cov(R_A,R_B) = 0.0058$.

Answer:

First, it is necessary to convert the variances to standard deviations.

$\sigma(R_A) = (0.0028)^{1/2} = 0.0529$

$\sigma(R_B) = (0.0124)^{1/2} = 0.1114$

Now, the correlation between the returns of Stock A and Stock B can be computed as follows:

$$Corr(R_A, R_B) = \frac{0.0058}{(0.0529)(0.1114)} = 0.9842$$

Coskewness and Cokurtosis

LO 16.l: Explain how coskewness and cokurtosis are related to skewness and kurtosis.

Previously, the first and second moments of mean and variance were applied to pairs of random variables. We can also apply techniques to identify the third and fourth moments for pairs of random variables that are similar to the measurements of skewness and kurtosis for individual variables. The third cross central moment is known as **coskewness** and the fourth cross central moment is known as **cokurtosis**.

There are p − 1 measures required for computed the *p*th moment. Figure 16.3 summarizes the number of measures required for each cross moment.

Figure 16.3: Cross Moment Measurements

Cross Moment	Number of Measurements
1st	0 cross means
2nd	1 covariance (cross variance)
3rd	2 coskewness (cross skewness)
4th	3 cokurtosis (cross kurtosis)

Dividing by the variance of one variable and the standard deviation of the other variable standardizes the cross third moment. The two coskewness measures are computed as:

$$s(X,X,Y) = \frac{E[(X - E[X])^2(Y - E[Y])]}{\sigma_X^2 \sigma_Y}$$

$$s(X,Y,Y) = \frac{E[(X - E[X])(Y - E[Y])^2]}{\sigma_X \sigma_Y^2}$$

Coskewness measures the likelihood of large directional movements occurring for one variable when the other variable is large. Coskewness measures are zero when there is no relationship between the sign of one variable when large moves occur with the other variable. Coskewness is always zero in a bivariate normal sample because the data is symmetrical and normally distributed.

We can estimate coskewness by applying an expectation operator as follows:

$$\hat{s}(X,X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2(Y_i - \hat{\mu}_Y)}{\hat{\sigma}_X^2 \hat{\sigma}_Y}$$

The three cokurtosis measures are computed as:

$$k(X,X,Y,Y) = \frac{E[(X - E[X])^2(Y - E[Y])^2]}{\sigma_X^2 \sigma_Y^2}$$

$$k(X,X,X,Y) = \frac{E[(X - E[X])^3(Y - E[Y])]}{\sigma_X^3 \sigma_Y}$$

$$k(X,Y,Y,Y) = \frac{E[(X - E[X])(Y - E[Y])^3]}{\sigma_X \sigma_Y^3}$$

The cokurtosis is computed using combinations of powers that add to 4. Note that the first cokurtosis measurement $k(X,X,Y,Y)$ is for the symmetrical case where there are two measurements from each variable (2,2). The asymmetric configurations are (1,3) and (3,1) where one of the variables measures to the third power and the other to the first power.

The symmetrical case provides the sensitivity of the magnitude of one series to the magnitude of the other series. The cokurtosis measure will be large if both series are large in magnitude at the same time. The other two asymmetrical cases indicate the agreement of the return signs when the power 3 return is large in magnitude.

The cokurtosis of a bivariate normal depends on the correlation. Figure 16.4 illustrates the relationship between cokurtosis and correlation for normal data and the symmetric case, $k(X,X,Y,Y)$. Notice that the correlation ranges between −1 and +1 and the cokurtosis ranges between 1 and 3, with the smallest value of 1 occurring when the correlation is equal to zero. When the correlation is zero, the returns are uncorrelated

with one another because both random variables are normally distributed. The cokurtosis then goes up symmetrically the further the correlation is away from zero.

Figure 16.4: Cokurtosis and Correlation for Symmetric Case

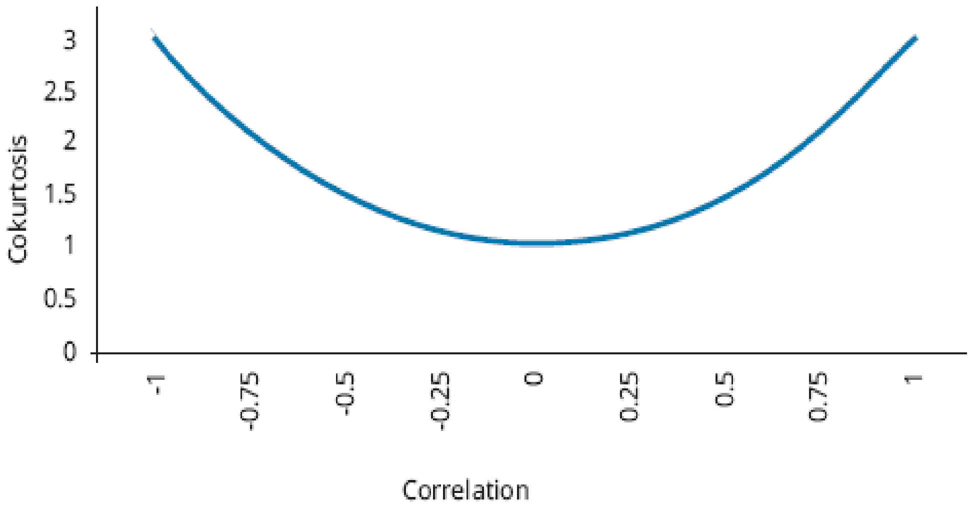
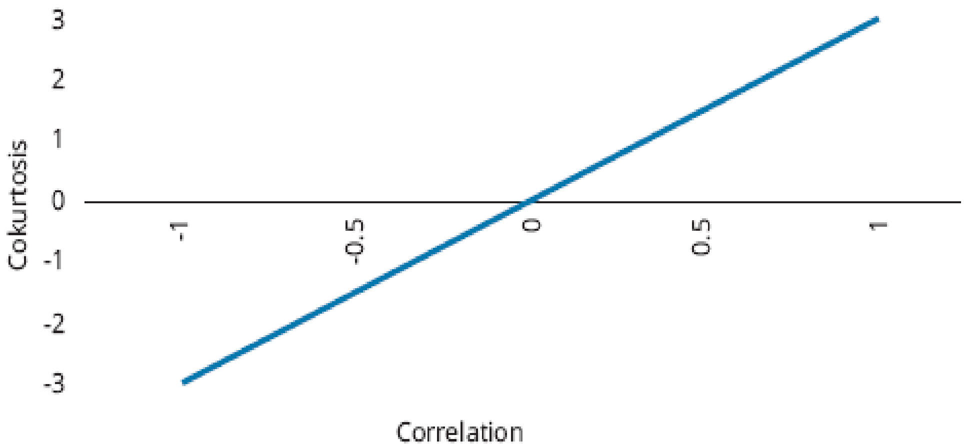


Figure 16.5 illustrates the relationship between cokurtosis and correlation for normal data and the asymmetrical cases where one series is to the power of three and the other is to the first power. The asymmetric cokurtosis ranges from −3 to +3 and is a linear relationship that is upward sloping as the correlation increases from −1 to +1.

Figure 16.5: Cokurtosis and Correlation for Asymmetric Cases



MODULE QUIZ 16.2

1. A junior analyst is assigned to estimate the first and second moments for an investment. Sample data was gathered that is assumed to represent the random data of the true population. Which of the following statements best describe the assumptions that are required to apply the central limit theorem (CLT) in estimating moments of this data set?
   A. Only the variance is finite.
   B. Both the mean and variance are finite.
   C. The random variables are normally distributed.
   D. The mean is finite and the random variables are normally distributed.
2. A distribution of returns that has a greater percentage of extremely large deviations from the mean:
   A. is positively skewed.
   B. is a symmetric distribution.
   C. has positive excess kurtosis.

D. has negative excess kurtosis.

3. The correlation of returns between Stocks A and B is 0.50. The covariance between these two securities is 0.0043, and the standard deviation of the return of Stock B is 26%. The variance of returns for Stock A is:
   A. 0.0331.
   B. 0.0011.
   C. 0.2656.
   D. 0.0112.

4. Consider the following probability matrix:

Probability Matrix			
Returns	R_B = 50%	R_B = 20%	R_B = –30%
R_A = –10%	40%	0%	0%
R_A = 10%	0%	30%	0%
R_A = 30%	0%	0%	30%

   The covariance between Stock A and B is closest to:
   A. −0.160.
   B. −0.055.
   C. 0.004.
   D. 0.020.

5. An analyst is graphing the cokurtosis and correlation for a pair of bivariate random variables that are normally distributed. For the symmetrical case of the three cokurtosis measures, $k(X,X,Y,Y)$, cokurtosis is graphed on the y-axis and correlation is graphed on the x-axis between –1 and +1. The shape of this graph should be best described as:
   A. an upward linear graph ranging in cokurtosis values between –3 and +3.
   B. a downward linear graph ranging in cokurtosis values between –1 and +1.
   C. a symmetrical curved graph with the maximum cokurtosis of 3 when the correlation is 0.
   D. a symmetrical curved graph with the minimum cokurtosis of 1 when the correlation is 0.

KEY CONCEPTS

LO 16.a

The sample mean, $\hat{\mu}$, and sample variance, $\hat{\sigma}^2$, for a sample of n independent and identically distributed (i.i.d.) random variables X_i are computed as:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2$$

LO 16.b

The sample mean is an estimator based on a known data set where all data points are observable. It is only an estimate of the true population mean.

LO 16.c

Point estimates are single (sample) values used to estimate population parameters, and the formula used to compute a point estimate is known as an estimator.

LO 16.d

The bias of an estimator measures the difference between the expected value of the estimator and the true population value:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

LO 16.e

The best linear unbiased estimator (BLUE) is the best estimator of the population mean available because it has the minimum variance of any linear unbiased estimator.

LO 16.f

A consistent estimator is one that as the sample size increases, the finite sample bias is reduced to zero and the variance of the estimator approaches zero.

LO 16.g

The law of large numbers (LLN) implies estimators converge to the true population value or where an average of many samples converges to the expected estimator. The central limit theorem (CLT) states that when the sample size is large, the sums of i.i.d. random variables will be normally distributed.

LO 16.h

Skewness is the third central moment of a distribution and refers to the extent to which the distribution of data is not symmetric around its mean. Kurtosis is the fourth central moment of a distribution and refers to how fat or thin the tails are in the distribution of data.

LO 16.i

The median calculation with an odd number sample size is:

$$\text{Median}(x) = x_{(n+1)/2}$$

The median calculation with an even number sample size is:

$$\text{Median}(x) = (1/2)(x_{n/2} + x_{n/2+1})$$

LO 16.j

The formulas for estimating the means of two random variables, X_i and Y_i are:

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

For i.i.d. data, we can apply the CLT and examine the joint behavior by stacking the two mean estimators into a vector:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{bmatrix}$$

LO 16.k

Covariance measures the extent to which two random variables tend to be above and below their respective means for each joint realization. It can be calculated as:

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

Correlation is a standardized measure of association between two random variables; it ranges in value from −1 to +1 and is equal to:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

LO 16.l

Coskewness measures the likelihood of large directional movements occurring for one variable when the other variable is large. Coskewness is zero when there is no relationship between the sign of one variable when large moves occur with the other variable.

The cokurtosis of a bivariate normal depends on the correlation. Cokurtosis for the symmetric case, k(X,X,Y,Y), ranges between +1 and +3, with the smallest value of 1 occurring when the correlation is equal to zero and the cokurtosis increases as the correlation moves away from zero. Cokurtosis for the asymmetrical cases range from −3 to +3 and is a linear relationship that is upward sloping as the correlation increases from −1 to +1.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 16.1

1. **D** The calculations for the sample mean and sample variance are shown in the following table:

X_i	Mean	Deviation	Squared Deviation
0.12	0.12	0.00	0.0000
0.25	0.12	0.13	0.0169
–0.01	0.12	–0.13	0.0169
0.36			0.0338

   The sum of all observations of returns for the asset is 0.36. Dividing this by the number of observations, 3, results in an unbiased estimate of the mean of 0.12. The third column subtracts the mean from the actual return for each year. The last column squares these deviations from the mean. The sum of the squared deviations is equal to 0.338 and dividing this by 2, for an unbiased estimate (n – 1) instead of the number of observations, results in an estimated variance of 0.0169. The standard deviation is then 0.13 (computed as the square root of the variance). (LO 16.a)

2. **B** The sample mean is an unbiased estimator of the population mean, because the expected value of the sample mean is equal to the population mean. The best linear unbiased estimator (BLUE) is the best estimator of the population mean available because it has the minimum variance of any linear unbiased estimator. (LO 16.e)

Module Quiz 16.2

1. **B** The CLT requires that the mean and variance are finite. The CLT does not require assumptions about the distribution of the random variables of the population. (LO 16.g)

2. **C** A distribution that has a greater percentage of extremely large deviations from the mean will be leptokurtic and will exhibit excess kurtosis (positive). The distribution will have fatter tails than a normal distribution. (LO 16.h)

3. **B** $\text{Corr}(R_A, R_B) = \dfrac{\text{Cov}(R_A, R_B)}{[\sigma(R_A)][\sigma(R_B)]}$

$$\sigma^2(R_A) = \left[\frac{\text{Cov}(R_A, R_B)}{[\sigma(R_B)]\text{Corr}(R_A, R_B)}\right]^2 = \left[\frac{0.0043}{(0.26)(0.5)}\right]^2 = 0.0331^2 = 0.0011$$

(LO 16.k)

4. **B** $\text{Cov}(R_A,R_B) = 0.4(-0.1 - 0.08)(0.5 - 0.17) + 0.3(0.1 - 0.08)(0.2 - 0.17) + 0.3(0.3 - 0.08)(-0.3 - 0.17) = -0.0546$

(LO 16.k)

5. **D** A symmetrical curved graph with the minimum cokurtosis of 1 when the correlation is 0. The graph will be an upward sloping linear relationship for the other two asymmetric cases of cokurtosis $k(X,Y,Y,Y)$ and $k(X,X,X,Y)$ (LO 16.l)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 6.

READING 17

HYPOTHESIS TESTING

Study Session 5

EXAM FOCUS

This reading provides insight into how risk managers make portfolio decisions on the basis of statistical analysis of samples of investment returns or other random economic and financial variables. We first focus on hypothesis testing procedures used to conduct tests concerned with population means and population variances. Specific tests reviewed include the *z*-test and the *t*-test. For the exam, you should be able to construct and interpret a confidence interval and know when and how to apply each of the test statistics discussed when conducting hypothesis testing.

MODULE 17.1: HYPOTHESIS TESTING BASICS

LO 17.a: Construct an appropriate null hypothesis and alternative hypothesis and distinguish between the two.

Hypothesis testing is the statistical assessment of a statement or idea regarding a population. For instance, a statement could be, "The mean return for the U.S. equity market is greater than zero." Given the relevant returns data, hypothesis testing procedures can be employed to test the validity of this statement at a given significance level.

A **hypothesis** is a statement about the value of a population parameter developed for the purpose of testing a theory or belief. Hypotheses are stated in terms of the population parameter to be tested, like the population mean, μ. For example, a researcher may be interested in the mean daily return on stock options. Hence, the hypothesis may be that the mean daily return on a portfolio of stock options is positive.

Hypothesis testing procedures, based on sample statistics and probability theory, are used to determine whether a hypothesis is a reasonable statement and should not be rejected or if it is an unreasonable statement and should be rejected. Any hypothesis test has six components:

- The null hypothesis, which specifies a value of the population parameter that is assumed to be true.

- The alternative hypothesis, which specifies the values of the test statistic over which we should reject the null.
- The test statistic, which is calculated from the sample data.
- The size of the test (commonly referred to as the significance level), which specifies the probability of rejecting the null hypothesis when it is true.
- The critical value, which is the value that is compared to the value of the test statistic to determine whether or not the null hypothesis should be rejected.
- The decision rule, which is the rule for deciding whether or not to reject the null hypothesis based on a comparison of the test statistic and the critical value.

PROFESSOR'S NOTE

Throughout this reading we use the more commonly used term *significance level* rather the *test size*. However, on the exam, recognize that if you see test size, it simply means significance level.

The Null Hypothesis and Alternative Hypothesis

The **null hypothesis**, designated H_0, is the hypothesis the researcher wants to reject. It is the hypothesis that is actually tested and is the basis for the selection of the test statistics. The null is generally a simple statement about a population parameter. Typical statements of the null hypothesis for the population mean include H_0: $\mu = \mu_0$, H_0: $\mu \leq \mu_0$, and H_0: $\mu \geq \mu_0$, where μ is the population mean and μ_0 is the hypothesized value of the population mean.

PROFESSOR'S NOTE

The null hypothesis always includes the *equal to* condition.

The **alternative hypothesis**, designated H_A, is what is concluded if there is sufficient evidence to reject the null hypothesis. It is usually the alternative hypothesis the researcher is really trying to assess. Why? Because you can never really prove anything with statistics, when the null hypothesis is discredited, the implication is that the alternative hypothesis is valid.

The Choice of the Null and Alternative Hypotheses

The most common null hypothesis will be an equal to hypothesis. The alternative is often the hoped-for hypothesis. When the null is that a coefficient is equal to zero, we hope to reject it and show the significance of the relationship.

When the null is less than or equal to, the (mutually exclusive) alternative is framed as greater than. If we are trying to demonstrate that a return is greater than the risk-free rate, this would be the correct formulation. We will have set up the null and alternative hypothesis so rejection of the null will lead to acceptance of the alternative, our goal in performing the test.

Hypothesis testing involves two statistics: the *test statistic* calculated from the sample data and the *critical value* of the test statistic. The value of the computed test statistic relative to the critical value is a key step in assessing the validity of a hypothesis.

A test statistic is calculated by comparing the point estimate of the population parameter with the hypothesized value of the parameter (i.e., the value specified in the null hypothesis). With reference to our option return example, this means we are concerned with the difference between the mean return of the sample and the hypothesized mean return. As indicated in the following expression, the test statistic is the difference between the sample statistic and the hypothesized value, scaled by the standard error of the sample statistic.

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the sample statistic}}$$

The standard error of the sample statistic is the adjusted standard deviation of the sample. When the sample statistic is the sample mean, $\bar{x}$, the standard error of the sample statistic for sample size *n*, is calculated as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

when the population standard deviation, σ, *is known*, or

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

when the population standard deviation, σ, is not known. In this case, it is estimated using the standard deviation of the sample, *s*.

PROFESSOR'S NOTE

Don't be confused by the notation here. A lot of the literature you will encounter in your studies simply uses the term $\sigma_{\bar{x}}$ for the standard error of the test statistic, regardless of whether the population standard deviation or sample standard deviation was used in its computation.

One-Tailed and Two-Tailed Tests of Hypotheses

LO 17.b: Differentiate between a one-sided and a two-sided test and identify when to use each test.

The alternative hypothesis can be one-sided or two-sided. A one-sided test is referred to as a **one-tailed test**, and a two-sided test is referred to as a **two-tailed test**. Whether the test is one- or two-sided depends on the proposition being tested. If a researcher wants to test whether the return on stock options is greater than zero, a one-tailed test should be used. However, a two-tailed test should be used if the research question is whether the return on options is simply different from zero. Two-sided tests allow for deviation on both sides of the hypothesized value (zero). In practice, most hypothesis tests are constructed as two-tailed tests.

A two-tailed test for the population mean may be structured as:

H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$

Because the alternative hypothesis allows for values above and below the hypothesized parameter, a two-tailed test uses two critical values (or rejection points).

The general decision rule for a two-tailed test is:

Reject H_0 if: test statistic > upper critical value or
test statistic < lower critical value

Let's look at the development of the decision rule for a two-tailed test using a *z*-distributed test statistic (a *z*-test) at a 5% level of significance, α = 0.05.

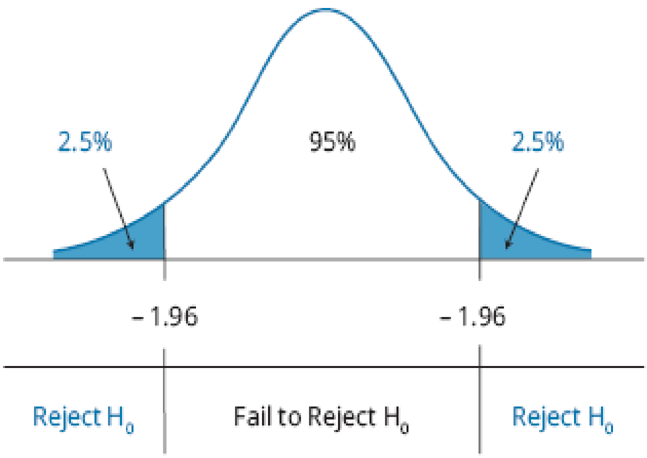
- At α = 0.05, the computed test statistic is compared with the critical z-values of ±1.96. The values of ±1.96 correspond to $\pm z_{\alpha/2} = \pm z_{0.025}$, which is the range of *z*-values within which 95% of the probability lies. These values are obtained from the cumulative probability table for the standard normal distribution (*z*-table), which is included at the back of this book.
- If the computed test statistic falls outside the range of critical *z*-values (i.e., test statistic > 1.96, or test statistic < –1.96), we reject the null and conclude that the sample statistic is sufficiently different from the hypothesized value.
- If the computed test statistic falls within the range ±1.96, we conclude that the sample statistic is not sufficiently different from the hypothesized value ($\mu = \mu_0$ in this case), and we fail to reject the null hypothesis.

The decision rule (rejection rule) for a two-tailed *z*-test at α = 0.05 can be stated as:

Reject H_0 if: test statistic < –1.96 or
test statistic > 1.96

Figure 17.1 shows the standard normal distribution for a two-tailed hypothesis test using the *z*-distribution. Notice that the significance level of 0.05 means that there is 0.05 / 2 = 0.025 probability (area) under each tail of the distribution beyond ±1.96.

Figure 17.1: Two-Tailed Hypothesis Test



EXAMPLE: Two-tailed test

A researcher has gathered data on the daily returns on a portfolio of call options over a recent 250-day period. The mean daily return has been 0.1%, and the sample standard deviation of daily portfolio returns is 0.25%. The researcher believes the

mean daily portfolio return is not equal to zero. **Construct** a hypothesis test of the researcher's belief.

Answer:

First, we need to specify the null and alternative hypotheses. The null hypothesis is the one the researcher expects to reject.

H_0: $\mu_0 = 0$ versus H_A: $\mu_0 \neq 0$

Because the null hypothesis is an equality, this is a two-tailed test. At a 5% level of significance, the critical *z*-values for a two-tailed test are ±1.96, so the decision rule can be stated as:

Reject H_0 if: test statistic < −1.96 or test statistic > +1.96

The standard error of the sample mean is the adjusted standard deviation of the sample. When the sample statistic is the sample mean, *x*, the standard error of the sample statistic for sample size *n* is calculated as:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Because our sample statistic here is a sample mean, the standard error of the sample mean for a sample size of 250 is $\frac{0.0025}{\sqrt{250}}$ and our test statistic is:

$$\frac{0.001}{\left(\frac{0.0025}{\sqrt{250}}\right)} = \frac{0.001}{0.000158} = 6.33$$

Because 6.33 > 1.96, we reject the null hypothesis that the mean daily option return is equal to zero. Note that when we reject the null, we conclude that the sample value is significantly different from the hypothesized value. We are saying that the two values are different from one another *after considering the variation in the sample*. That is, the mean daily return of 0.001 is statistically different from zero given the sample's standard deviation and size.

For a one-tailed hypothesis test of the population mean, the null and alternative hypotheses are either:

- upper tail: H_0: $\mu \leq \mu_0$ versus H_A: $\mu > \mu_0$, or
- lower tail: H_0: $\mu \geq \mu_0$ versus H_A: $\mu < \mu_0$.

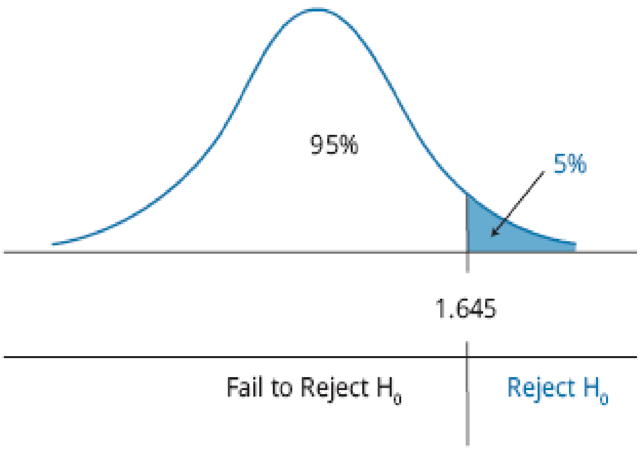
The appropriate set of hypotheses depends on whether we believe the population mean, μ, to be greater than (upper tail) or less than (lower tail) the hypothesized value, μ_0. Using a *z*-test at the 5% level of significance, the computed test statistic is compared with the critical values of 1.645 for the upper tail tests (i.e., H_A: $\mu > \mu_0$) or −1.645 for lower tail tests (i.e., H_A: $\mu < \mu_0$). These critical values are obtained from a *z*-table, where $-z_{0.05} = -1.645$ corresponds to a cumulative probability equal to 5%, and the $z_{0.05}$ = 1.645 corresponds to a cumulative probability of 95% (1 − 0.05).

Let's use the upper tail test structure where H_0: $\mu \leq \mu_0$ and H_A: $\mu > \mu_0$.

- If the calculated test statistic is greater than 1.645, we conclude that the sample statistic is sufficiently greater than the hypothesized value. In other words, we reject the null hypothesis.
- If the calculated test statistic is less than 1.645, we conclude that the sample statistic is not sufficiently different from the hypothesized value, and we fail to reject the null hypothesis.

Figure 17.2 shows the standard normal distribution and the rejection region for a one-tailed test (upper tail) at the 5% level of significance.

Figure 17.2: One-Tailed Hypothesis Test



> EXAMPLE: **One-tailed test**
>
> **Perform** a *z*-test using the option portfolio data from the previous example to test the belief that option returns are positive.
>
> **Answer:**
>
> In this case, we use a one-tailed test with the following structure:
>
> H_0: $\mu \leq 0$ versus H_A: $\mu > 0$
>
> The appropriate decision rule for this one-tailed *z*-test at a significance level of 5% is:
>
> Reject H_0 if: test statistic > 1.645
>
> The test statistic is computed the same way, regardless of whether we are using a one-tailed or two-tailed test. From the previous example, we know the test statistic for the option return sample is 6.33. Because 6.33 > 1.645, we reject the null hypothesis and conclude that mean returns are statistically greater than zero at a 5% level of significance.

Type I and Type II Errors

LO 17.c: Explain the difference between Type I and Type II errors and how these relate to the size and power of a test.

Keep in mind that hypothesis testing is used to make inferences about the parameters of a given population on the basis of statistics computed for a sample that is drawn from that population. We must be aware that there is some probability that the sample, in some way, does not represent the population and any conclusion based on the sample about the population may be made in error.

When drawing inferences from a hypothesis test, there are two types of errors:

- **Type I error**: the rejection of the null hypothesis when it is actually true.
- **Type II error**: the failure to reject the null hypothesis when it is actually false.

The significance level is the probability of making a Type I error (rejecting the null when it is true) and is designated by the Greek letter alpha (α). For instance, a significance level of 5% ($\alpha = 0.05$) means there is a 5% chance of rejecting a true null hypothesis. When conducting hypothesis tests, a significance level must be specified in order to identify the critical values needed to evaluate the test statistic.

The decision for a hypothesis test is to either reject the null hypothesis or fail to reject the null hypothesis. Note that it is statistically incorrect to say "accept" the null hypothesis; it can only be supported or rejected. The decision rule for rejecting or failing to reject the null hypothesis is based on the distribution of the test statistic. For example, if the test statistic follows a normal distribution, the decision rule is based on critical values determined from the standard normal distribution (z-distribution). Regardless of the appropriate distribution, it must be determined if a one-tailed or two-tailed hypothesis test is appropriate before a decision rule (rejection rule) can be determined.

A decision rule is specific and quantitative. Once we have determined whether a one- or two-tailed test is appropriate, the significance level we require, and the distribution of the test statistic, we can calculate the exact critical value for the test statistic. Then we have a decision rule of the following form: if the test statistic is (greater, less than) the value X, reject the null.

While the significance level of a test is the probability of rejecting the null hypothesis when it is true, the **power of a test** is the probability of correctly rejecting the null hypothesis when it is false. The power of a test is actually one minus the probability of making a Type II error, or 1 − P(Type II error). In other words, the probability of rejecting the null when it is false (power of the test) equals one minus the probability of *not* rejecting the null when it is false (Type II error). When more than one test statistic may be used, the power of the test for the competing test statistics may be useful in deciding which test statistic to use. Ordinarily, we wish to use the test statistic that provides the most powerful test among all possible tests.

Figure 17.3 shows the relationship between the level of significance, the power of a test, and the two types of errors.

Figure 17.3: Type I and Type II Errors in Hypothesis Testing

		True Condition	
		H_0 is true	H_0 is false
Decision	Do not reject H_0	Correct decision	Incorrect decision Type II error
	Reject H_0	Incorrect decision Type I error Significance level, α, = P(Type I error)	Correct decision Power of the test = 1 – P(Type II error)

Sample size and the choice of significance level (Type I error probability) will together determine the probability of a Type II error. The relation is not simple, however, and calculating the probability of a Type II error in practice is quite difficult. Decreasing the significance level (probability of a Type I error) from 5% to 1%, for example, will increase the probability of failing to reject a false null (Type II error) and, therefore, reduce the power of the test. Conversely, for a given sample size, we can increase the power of a test only with the cost that the probability of rejecting a true null (Type I error) increases. For a given significance level, we can decrease the probability of a Type II error and increase the power of a test, only by increasing the sample size.

The Relation Between Confidence Intervals and Hypothesis Tests

LO 17.d: Understand how a hypothesis test and a confidence interval are related.

A **confidence interval** is a range of values within which the researcher believes the true population parameter may lie.

A confidence interval is determined as:

$$\left\{\left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} - \begin{pmatrix}\text{critical}\\\text{value}\end{pmatrix}\begin{pmatrix}\text{standard}\\\text{error}\end{pmatrix}\right] \leq \begin{matrix}\text{population}\\\text{parameter}\end{matrix} \leq \left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} + \begin{pmatrix}\text{critical}\\\text{value}\end{pmatrix}\begin{pmatrix}\text{standard}\\\text{error}\end{pmatrix}\right]\right\}$$

The interpretation of a confidence interval is that for a level of confidence of 95%, for example, there is a 95% probability that the true population parameter is contained in the interval.

From the previous expression, we see that a confidence interval and a hypothesis test are linked by the critical value. For example, a 95% confidence interval uses a critical value associated with a given distribution at the 5% level of significance. Similarly, a hypothesis test would compare a test statistic to a critical value at the 5% level of significance. To see this relationship more clearly, the expression for the confidence interval can be manipulated and restated as:

$$-\text{critical value} \leq \text{test statistic} \leq +\text{critical value}$$

This is the range within which we fail to reject the null for a two-tailed hypothesis test at a given level of significance.

EXAMPLE: Confidence interval

Using option portfolio data from the previous examples, **construct** a 95% confidence interval for the population mean daily return over the 250-day sample period. Use a z-distribution. **Decide** if the hypothesis μ = 0 should be rejected.

Answer:

Given a sample size of 250 with a standard deviation of 0.25%, the standard error can be computed as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.25}{\sqrt{250}} = 0.0158\%$$

At the 5% level of significance, the critical z-values for the confidence interval are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$. Thus, given a sample mean equal to 0.1%, the 95% confidence interval for the population mean is:

$0.1 - 1.96(0.0158) \leq \mu \leq 0.1 + 1.96(0.0158)$, or

$0.069\% \leq \mu \leq 0.1310\%$

Because there is a 95% probability that the true mean is within this confidence interval, we can reject the hypothesis μ = 0 because 0 is not within the confidence interval.

Notice the similarity of this analysis with our test of whether μ = 0. We rejected the hypothesis μ = 0 because the sample mean of 0.1% is more than 1.96 standard errors from zero. Based on the 95% confidence interval, we reject μ = 0 because zero is more than 1.96 standard errors from the sample mean of 0.1%.

Statistical Significance vs. Practical Significance

Statistical significance does not necessarily imply practical significance. For example, we may have tested a null hypothesis that a strategy of going long all the stocks that satisfy some criteria and shorting all the stocks that do not satisfy the criteria resulted in returns that were less than or equal to zero over a 20-year period. Assume we have rejected the null in favor of the alternative hypothesis that the returns to the strategy are greater than zero (positive). This does not necessarily mean that investing in that strategy will result in economically meaningful positive returns. Several factors must be considered.

One important consideration is transactions costs. Once we consider the costs of buying and selling the securities, we may find that the mean positive returns to the strategy are not enough to generate positive returns. Taxes are another factor that may make a seemingly attractive strategy a poor one in practice. A third reason that statistically significant results may not be economically significant is risk. In the strategy just discussed, we have additional risk from short sales (they may have to be closed out earlier than in the test strategy). Because the statistically significant results were for a period of 20 years, it may be the case that there is significant variation from year to year in the returns from the strategy, even though the mean strategy return is greater than zero. This variation in returns from period to period is an additional risk to the strategy that is not accounted for in our test of statistical significance.

Any of these factors could make committing funds to a strategy unattractive, even though the statistical evidence of positive returns is highly significant. By the nature of statistical tests, a very large sample size can result in highly (statistically) significant results that are quite small in absolute terms.

MODULE QUIZ 17.1

1. Austin Roberts believes the mean price of houses in the area is greater than \$145,000. A random sample of 36 houses in the area has a mean price of \$149,750. The population standard deviation is \$24,000, and Roberts wants to conduct a hypothesis test at a 1% level of significance.The appropriate alternative hypothesis is:
   A. H_A: μ < \$145,000.
   B. H_A: μ ± \$145,000.
   C. H_A: μ ≥ \$145,000.
   D. H_A: μ > \$145,000.

2. Which of the following statements about hypothesis testing is most accurate?
   A. The power of a test is one minus the probability of a Type I error.
   B. The probability of a Type I error is equal to the significance level of the test.
   C. To test the claim that X is greater than zero, the null hypothesis would be H_0: X > 0.
   D. If you can disprove the null hypothesis, then you have proven the alternative hypothesis.

MODULE 17.2: HYPOTHESIS TESTING RESULTS

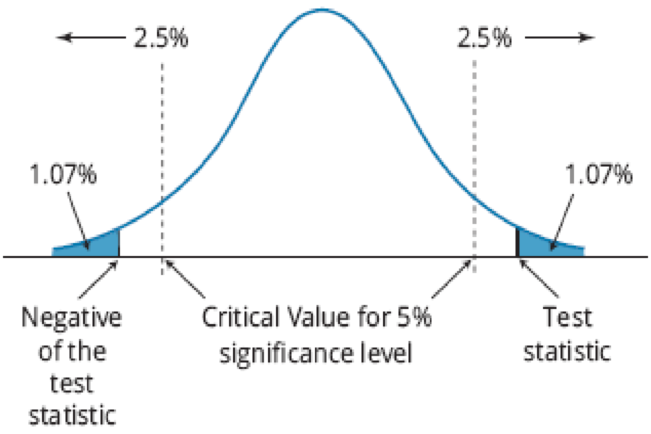
LO 17.e: Explain what the *p*-value of a hypothesis test measures.

The *p*-Value

The ***p*-value** is the probability of obtaining a test statistic that would lead to a rejection of the null hypothesis, assuming the null hypothesis is true. It is the smallest level of significance for which the null hypothesis can be rejected. For one-tailed tests, the *p*-value is the probability that lies above the computed test statistic for upper tail tests or below the computed test statistic for lower tail tests. For two-tailed tests, the *p*-value is the probability that lies above the positive value of the computed test statistic *plus* the probability that lies below the negative value of the computed test statistic.

Consider a two-tailed hypothesis test about the mean value of a random variable at the 95% significance level where the test statistic is 2.3, greater than the upper critical value of 1.96. If we consult the *z*-table, we find the probability of getting a value greater than 2.3 is (1 − 0.9893) = 1.07%. Because it's a two-tailed test, our *p*-value is 2 × 1.07 = 2.14%, as illustrated in Figure 17.4. At a 3%, 4%, or 5% significance level, we would reject the null hypothesis, but at a 2% or 1% significance level, we would not. Many researchers report *p*-values without selecting a significance level and allow the reader to judge how strong the evidence for rejection is.

Figure 17.4: Two-Tailed Hypothesis Test With *p*-Value = 2.14%



Confidence Intervals for Hypothesis Tests

LO 17.f: Construct and apply confidence intervals for one-sided and two-sided hypothesis tests and interpret the results of hypothesis tests with a specific confidence level.

As mentioned earlier, confidence interval estimates result in a range of values within which the actual value of a parameter will lie, given the probability of 1 − α, where alpha, α, is the level of significance for the confidence interval, and the probability 1 − α is the degree of confidence. Recall the confidence interval for a two-tailed test:

$$\left\{\left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} - \begin{pmatrix}\text{critical}\\\text{value}\end{pmatrix}\begin{pmatrix}\text{standard}\\\text{error}\end{pmatrix}\right] \le \begin{matrix}\text{population}\\\text{parameter}\end{matrix} \le \left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} + \begin{pmatrix}\text{critical}\\\text{value}\end{pmatrix}\begin{pmatrix}\text{standard}\\\text{error}\end{pmatrix}\right]\right\}$$

Confidence intervals can also be provided for one-tailed tests as either:

Upper tail:

$$\left\{\left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} - \begin{pmatrix}\text{critical}\\\text{value}\end{pmatrix}\begin{pmatrix}\text{standard}\\\text{error}\end{pmatrix}\right] \le \begin{matrix}\text{population}\\\text{parameter}\end{matrix} \le \infty\right\}$$

Lower tail:

$$\left\{-\infty \le \begin{matrix}\text{population}\\\text{parameter}\end{matrix} \le \left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} + \begin{pmatrix}\text{critical}\\\text{value}\end{pmatrix}\begin{pmatrix}\text{standard}\\\text{error}\end{pmatrix}\right]\right\}$$

With hypothesis testing, the choice between using a critical value based on the *t*-distribution or the *z*-distribution depends on the sample size, the distribution of the population, and whether the variance of the population is known or unknown.

The *t*-Test

The ***t*-test** is a widely used hypothesis test that employs a test statistic that is distributed according to a *t*-distribution. Following are the rules for when it is appropriate to use the *t*-test for hypothesis tests of the population mean.

Use the *t*-test if the population variance is unknown and either of the following conditions exist:

- The sample is large ($n \geq 30$).
- The sample is small ($n < 30$), but the distribution of the population is normal or approximately normal.

If the sample is small and the distribution is nonnormal, we have no reliable statistical test.

The computed value for the test statistic based on the *t*-distribution is referred to as the *t*-statistic. For hypothesis tests of a population mean, a *t*-statistic with n − 1 degrees of freedom is computed as:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where:
$\bar{x}$ = sample mean
μ_0 = hypothesized population mean (i.e., the null)
s = standard deviation of the sample
n = sample size

PROFESSOR'S NOTE

This computation is not new. It is the same test statistic computation that we have been performing all along. Note the use of the sample standard deviation, s, in the standard error term in the denominator.

To conduct a *t*-test, the *t*-statistic is compared to a critical *t*-value at the desired level of significance with the appropriate degrees of freedom.

In the real world, the underlying variance of the population is rarely known, so the *t*-test enjoys widespread application.

The *z*-Test

The ***z*-test** is the appropriate hypothesis test of the population mean when the *population is normally distributed with known variance*. The computed test statistic used with the *z*-test is referred to as the *z*-statistic. The *z*-statistic for a hypothesis test for a population mean is computed as follows:

$$z\text{-statistic} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

where:
$\bar{x}$ = sample mean
μ_0 = hypothesized population mean
σ = standard deviation of the population
n = sample size

To test a hypothesis, the *z*-statistic is compared to the critical *z*-value corresponding to the significance of the test. Critical *z*-values for the most common levels of significance are displayed in Figure 17.5. You should memorize these critical values for the exam.

Figure 17.5: Critical *z*-Values

Level of Significance	Two-Tailed Test	One-Tailed Test
0.10 = 10%	±1.65	+1.28 or −1.28
0.05 = 5%	±1.96	+1.65 or −1.65
0.01 = 1%	±2.58	+2.33 or −2.33

When the *sample size is large* and the *population variance is unknown*, the *z*-statistic is:

$$\text{z-statistic} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where:
$\overline{x}$ = sample mean
μ_0 = hypothesized population mean
s = standard deviation of the sample
n = sample size

Note the use of the sample standard deviation, *s*, versus the population standard deviation, σ. Remember, this is acceptable if the sample size is large, although the *t*-statistic is the more conservative measure when the population variance is unknown.

EXAMPLE: *z*-test or *t*-test?

Referring to our previous option portfolio mean return problem once more, **determine** which test statistic (*z* or *t*) should be used and the difference in the likelihood of rejecting a true null with each distribution.

Answer:

The population variance for our sample of returns is unknown. Hence, the *t*-distribution is appropriate. With 250 observations, however, the sample is considered to be large, so the *z*-distribution would also be acceptable. This is a trick question—either distribution, *t* or *z*, is appropriate. With regard to the difference in the likelihood of rejecting a true null, because our sample is so large, the critical values for the *t* and *z* are almost identical. Hence, there is almost no difference in the likelihood of rejecting a true null.

EXAMPLE: The *z*-test

When your company's gizmo machine is working properly, the mean length of gizmos is 2.5 inches. However, from time to time the machine gets out of alignment and produces gizmos that are either too long or too short. When this happens, production is stopped and the machine is adjusted. To check the machine, the quality control department takes a gizmo sample each day. Today, a random sample of 49 gizmos showed a mean length of 2.49 inches. The population standard deviation is known to be 0.021 inches. Using a 5% significance level, **determine** if the machine should be shut down and adjusted.

Answer:

Let μ be the mean length of all gizmos made by this machine, and let $\bar{x}$ be the corresponding mean for the sample. A common hypothesis testing procedure is outlined as follows:

Statement of hypothesis. For the information provided, the null and alternative hypotheses are appropriately structured as:

H_0: μ = 2.5 (The machine does not need an adjustment.)

H_A: μ ≠ 2.5 (The machine needs an adjustment.)

Note that because this is a two-tailed test, H_A allows for values above and below 2.5.

Select the appropriate test statistic. Because the population variance is known and the sample size is > 30, the *z*-statistic is the appropriate test statistic. The *z*-statistic is computed as:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Specify the level of significance. The level of significance is given at 5%, implying that we are willing to accept a 5% probability of rejecting a true null hypothesis.

State the decision rule regarding the hypothesis. The ≠ sign in the alternative hypothesis indicates that the test is two-tailed with two rejection regions, one in each tail of the standard normal distribution curve. Because the total area of both rejection regions combined is 0.05 (the significance level), the area of the rejection region in each tail is 0.025. You should know that the critical *z*-values for $\pm z_{0.025}$ are ±1.96. This means that the null hypothesis should not be rejected if the computed z-statistic lies between −1.96 and +1.96 and should be rejected if it lies outside of these critical values. The decision rule can be stated as:

Reject H_0 if: z-statistic < $-z_{0.025}$ or z-statistic > $z_{0.025}$, or equivalently

Reject H_0 if: z-statistic < −1.96 or z-statistic > + 1.96

Collect the sample and calculate the test statistic. The value of $\bar{x}$ from the sample is 2.49. Because σ is given as 0.021, we calculate the *z*-statistic using $\bar{x}$ as follows:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{2.49 - 2.5}{0.021 / \sqrt{49}} = \frac{-0.01}{0.003} = -3.33$$

Make a decision regarding the hypothesis. The calculated value of the *z*-statistic is −3.33. Because this value is less than the critical value, $-z_{0.025}$ = −1.96, it falls in the rejection region in the left tail of the *z*-distribution. Hence, there is sufficient evidence to reject H_0.

Make a decision based on the results of the test. Based on the sample information and the results of the test, it is concluded that the machine is out of adjustment and should be shut down for repair.

Testing the Equality of Means

LO 17.g: Identify the steps to test a hypothesis about the difference between two population means.

In finance, we are often interested in testing whether the means of two populations are equal to each other. This is equivalent to testing whether the difference between the two means is zero.

If we assume two series (X and Y) are each independent and identically distributed (i.i.d.) and have a covariance of cov(X,Y), the appropriate test statistic is:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{s_X^2 + s_Y^2 - 2\text{cov}(X, Y)}{n}}}$$

This test statistic has a standard normal distribution when the null hypothesis is true.

The steps to test the hypothesis that the means are equal would then follow the standard hypothesis testing procedure. The null hypothesis would be that the difference between the two is equal to zero, versus the alternative that it is not equal to zero. Given the test size and the appropriate critical value, the null would be rejected or fail to be rejected by comparing the test statistic to the critical value.

Multiple Hypothesis Testing

LO 17.h: Explain the problem of multiple testing and how it can lead to biased results.

Multiple testing means testing multiple different hypothesis on the same data set. For example, suppose we are testing 10 active trading strategies against a buy-and-hold trading strategy. The problem is that if we keep testing different strategies against the same null hypothesis, it is highly likely we are eventually going to reject one of them. The problem with this is that the alpha (the probability of incorrectly rejecting a true null) is only accurate for one single hypothesis test. As we test more and more strategies, the actual alpha of this repeated testing grows larger, and as alpha grows larger, the probability of a Type I error increases.

MODULE QUIZ 17.2

1. The most likely bias to result from testing multiple hypotheses on a single data set is that the value of:
   A. a Type I error will increase.
   B. a Type II error will increase.
   C. the critical value will increase.
   D. the test statistic will increase.
2. Austin Roberts believes the mean price of houses in the area is greater than $145,000. A random sample of 36 houses in the area has a mean price of $149,750. The population standard deviation is $24,000, and Roberts wants to conduct a hypothesis test at a 1% level of significance.The value of the calculated test statistic is closest to:

A. $z = 0.67$.
B. $z = 1.19$.
C. $z = 4.00$.
D. $z = 8.13$.

KEY CONCEPTS

LO 17.a

The hypothesis testing process requires a statement of a null and an alternative hypothesis, the selection of the appropriate test statistic, specification of the significance level, a decision rule, the calculation of a sample statistic, a decision regarding the hypotheses based on the test, and a decision based on the test results.

The test statistic is the value that a decision about a hypothesis will be based on. For a test about the value of the mean of a distribution:

$$\text{test statistic} = \frac{\text{sample mean} - \text{hypothesized mean}}{\text{standard error of sample mean}}$$

LO 17.b

A two-tailed test results from a two-sided alternative hypothesis (e.g., H_A: $\mu \neq \mu_0$). A one-tailed test results from a one-sided alternative hypothesis (e.g., H_A: $\mu > \mu_0$, or H_A: $\mu < \mu_0$).

LO 17.c

		True Condition	
		H_0 is true	H_0 is false
Decision	Do not reject H_0	Correct decision	Incorrect decision Type II error
Decision	Reject H_0	Incorrect decision Type I error Significance level, α, = P(Type I error)	Correct decision Power of the test = 1 – P(Type II error)

LO 17.d

Hypothesis testing compares a computed test statistic to a critical value at a stated level of significance, which is the decision rule for the test.

A hypothesis about a population parameter is rejected when the sample statistic lies outside a confidence interval around the hypothesized value for the chosen level of significance.

LO 17.e

The *p*-value is the probability of obtaining a test statistic that would lead to a rejection of the null hypothesis, assuming the null hypothesis is true. It is the smallest level of significance for which the null hypothesis can be rejected.

LO 17.f

For hypothesis tests of a population mean, a *t*-statistic with n − 1 degrees of freedom is computed as:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

To conduct a *t*-test, the *t*-statistic is compared to a critical *t*-value at the desired level of significance with the appropriate degrees of freedom.

LO 17.g

The appropriate test statistic to test whether the means of two populations are equal to each other is:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2 + s_Y^2 - 2\text{cov}(X, Y)}{n}}}$$

This test statistic has a standard normal distribution when the null hypothesis is true.

LO 17.h

Multiple testing means testing multiple different hypothesis on the same data set. The problem with multiple testing is that the alpha is only accurate for one single hypothesis test. As we test more and more strategies, the alpha of this repeated testing grows larger, and as alpha grows larger, the probability of a Type I error increases.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 17.1

1. **D** H_A: μ > $145,000. (LO 17.b)

2. **B** The probability of getting a test statistic outside the critical value(s) when the null is true is the level of significance and is the probability of a Type I error. The power of a test is one minus the probability of a Type II error. Hypothesis testing does not prove a hypothesis; we either reject the null or fail to reject it. The appropriate null would be X ≤ 0 with X > 0 as the alternative hypothesis. (LO 17.c)

Module Quiz 17.2

1. **A** With multiple testing, the alpha (the probability of incorrectly rejecting a true null) is only accurate for one single hypothesis test. As we test more and more strategies, the actual alpha of this repeated testing grows larger, and as alpha grows larger, the probability of a Type I error increases. (LO 17.h)

2. **B** $z = \frac{149{,}750 - 145{,}000}{24{,}000 / \sqrt{36}} = 1.1875$

   (LO 17.f)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 7.

READING 18

LINEAR REGRESSION

Study Session 6

EXAM FOCUS

Linear regression refers to the process of representing relationships with linear equations where there is one dependent variable being explained by one or more independent variables. Typically, we estimate a regression equation using ordinary least squares (OLS), which minimizes the sum of squared errors in the sample data. For the exam, be able to conduct hypothesis tests, calculate confidence intervals, and remember the assumptions underlying the regression model. Finally, understand how to interpret a regression equation.

MODULE 18.1: REGRESSION ANALYSIS

LO 18.a: Describe the models that can be estimated using linear regression and differentiate them from those which cannot.

Regression analysis seeks to measure how changes in one variable, called a **dependent** (or **explained**) variable can be explained by changes in one or more other variables called the **independent** (or **explanatory**) variables. This relationship is captured by estimating a linear equation.

As an example, we want to capture the relationship between hedge fund returns and lockup periods.

For this simple two-variable case (i.e., one explained and one explanatory variable), the function is:

$$E(\text{return}) = \alpha + \beta \times (\text{lockup period})$$

Or more generally:

$$E(Y) = \alpha + \beta \times (X)$$

Which we can write as:

$$Y = \alpha + \beta \times (X) + \varepsilon$$

where:
β = the regression or slope coefficient; sensitivity of Y to changes in X

α = value of Y when X = 0
ε = random error or shock; unexplained (by X) component of Y

This error may be reduced by using more independent variables or by using different, more appropriate independent variables.

Note that the interpretation of α changes when X cannot realistically take on a value of 0. In such a case, α is the value that ensures that the mean of Y lies on the fitted regression line.

Linear Regression Conditions

To use linear regression, three conditions need to be satisfied:

1. The relationship between Y and X should be linear (discussed later).
2. The error term must be additive (i.e., the variance of the error term is independent of the observed data).
3. All X variables should be observable (i.e., makes the model inappropriate when you have missing data).

The term *linear* has implications for both the independent variable(s) and the unknown parameters (i.e., the coefficients). However, appropriate transformations of the independent variable(s) can make a nonlinear relationship amenable to be fitted using a linear model.

If the relationship between the dependent variable (Y) and an independent variable (X) is nonlinear, then an analyst would do that transformation first and then enter the transformed value into the linear equation as *X*. For example, in estimating a utility function as a function of consumption, we might allow for the property of diminishing marginal utility by transforming consumption into a logarithm of consumption. In other words, the actual relationship is:

$$E(\text{utility}) = \alpha + \beta \times \ln(\text{amount consumed})$$

Here we let Y = utility and X = ln (amount consumed) and then estimate: $E(Y) = \alpha + \beta \times (X)$ using linear techniques.

A second interpretation of the term *linear* applies to the unknown coefficients. It specifies that the dependent variable is a linear function of the coefficients. For example, consider an unknown parameter, *p*, in a function: $Y = \alpha + \beta X^p + \varepsilon$. In this instance, βX^p contains two unknown parameters (β and p) and p does not enter the model multiplicatively and, hence, it would not be appropriate to apply linear regression in such a case.

MODULE QUIZ 18.1

1. Generally, if the value of the independent variable is zero, then the expected value of the dependent variable would be equal to:
   A. the slope coefficient.
   B. the intercept coefficient.
   C. the error term.
   D. the residual.
2. The error term represents the portion of:

A. the dependent variable that is not explained by the independent variable(s) but could possibly be explained by adding additional independent variables.
B. the dependent variable that is explained by the independent variable(s).
C. the independent variables that are explained by the dependent variable.
D. the dependent variable that is explained by the error in the independent variable(s).

3. A linear regression function assumes that the relation being modeled must be linear in:
A. both the variables and the coefficients.
B. the coefficients but not necessarily the variables.
C. the variables but not necessarily the coefficients.
D. neither the variables nor the coefficients.

MODULE 18.2: ORDINARY LEAST SQUARES ESTIMATION

LO 18.b: Interpret the results of an ordinary least squares (OLS) regression with a single explanatory variable.

Ordinary least squares (OLS) estimation is a process that estimates the parameters α and β in an effort to minimize the squared residuals (i.e., error terms).

Rewriting our regression equation: $\varepsilon_i = Y_i - (\alpha + \beta \times X_i)$; the OLS sample coefficients are those that minimize $\sum \varepsilon_i^2 = \sum [Y_i - (\alpha + \beta \times X_i)]^2$

The estimated **slope coefficient** (β) for the regression line describes the change in *Y* for a one-unit change in *X*. The slope term is calculated as:

$$\beta = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

The **intercept** term (α) is the line's intersection with the *Y*-axis at X = 0. A property of the least squares method is that the intercept term may be expressed as:

$$\alpha = \overline{Y} - \beta\overline{X}$$

where:
$\overline{Y}$ = mean of Y
$\overline{X}$ = mean of X

The intercept equation highlights the fact that the regression line passes through a point with coordinates equal to the mean of the independent and dependent variables.

Interpreting Regression Results

The intercept term, α, is the value of the dependent variable when the independent variable is equal to zero. The slope coefficient, β, is the estimated change in the dependent variable for a one-unit change in that independent variable. In the case where the model uses multiple independent variables, the interpretation of the slope coefficient captures the change in the dependent variable for a one-unit change in the

independent variable, holding the other independent variables constant. As you will see in the next reading, this is why the slope coefficients in a multiple regression are sometimes called **partial slope coefficients**.

EXAMPLE: Regression model with one explanatory variable

The mean annual return $\overline{Y}$ over the past 20 years for a specific stock is 11%, while that for the market $\overline{X}$ is 8.4%. The covariance of annual returns for the stock and the market (σ_{XY}) and the variance of the market (σ_X^2) are shown in the following variance-covariance matrix:

$$\begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix} = \begin{pmatrix} 151.22 & 132.11 \\ 132.11 & 181.40 \end{pmatrix}$$

Calculate the estimated slope coefficient and intercept, and **interpret** the regression results.

Answer:

$$\hat{\beta} = \frac{\text{Cov(X,Y)}}{\text{Var X}} = \frac{\sigma_{X,Y}}{\sigma_X^2} = \frac{132.11}{181.40} = 0.73$$

$$\alpha = \overline{Y} - \beta\overline{X} = 0.11 - 0.73 \times 0.084 = 0.049$$

Interpretation of the coefficients:

β: A 1% increase in returns in the market would lead to an increase in 0.73% increase in the return on the stock.

α: If the market return is 0%, Stock A's return would be 0.049 or 4.9%.

Dummy Variables

Observations for most independent variables (e.g., firm size, level of GDP, and interest rates) can take on a wide range of values. However, there are occasions when the independent variable is binary in nature—it is either *on* or *off*. Independent variables that fall into this category are called **dummy variables** and are often used to quantify the impact of qualitative variables.

Dummy variables are assigned a value of 0 or 1. For example, in a time series regression of monthly stock returns, you could employ a January dummy variable that would take on the value of 1 if a stock return occurred in January, and 0 if it occurred in any other month. The purpose of including the January dummy variable would be to see if stock returns in January were significantly different than stock returns in all other months of the year.

Coefficient of Determination of a Regression (R^2)

LO 18.h: Estimate the correlation coefficient from the R^2 measure obtained in linear regressions with a single explanatory variable.

The R^2 of a regression model captures the fit of the model; it represents the proportion of variation in the dependent variable that is explained by the independent variable(s). For a regression model with a single independent variable, R^2 is the square of the correlation between the independent and dependent variable.

$$R^2 = r^2_{X,Y}$$

where $r_{X,Y}$ = correlation between X and Y

Assumptions Underlying Linear Regression

LO 18.c: Describe the key assumptions of OLS parameter estimation.

OLS regression requires a number of assumptions. Most of the major assumptions pertain to the regression model's residual term (i.e., the error term). The key assumptions are as follows:

- The expected value of the error term, conditional on the independent variable, is zero $[E(\varepsilon_i \mid X_i) = 0]$. This means that X has no information about the location of ε. This assumption is not directly testable; OLS estimates using sample data ensure that the shocks are always uncorrelated with Xs. Evaluation of whether this assumption is reasonable requires an examination of the data generating process. Generally, a violation would be evidenced by the following:
  - **Survivorship**, or **sample selection**, **bias**: Survivorship bias occurs when the observations are collected after-the-fact (e.g., companies that get dropped from an index are not included in the sample). Sample selection bias occurs when occurrence of an event (i.e., an observation) is contingent on specific outcomes. For example, mortgage refinancing is severely curtailed during falling housing prices and, hence, the sample of actual refinancing transactions is therefore more likely to occur during a rising home-price environment.
  - **Simultaneity bias**: This happens when the values of X and Y are simultaneously determined. For example, trading volume and volatility are related; volume increases during volatile times.
  - **Omitted variables**: Important explanatory (i.e., X) variables are not excluded from the model. If they are, the errors will capture the influence of the omitted variables. Omission of important variables cause the coefficients to be biased and may indicate nonexistent (i.e., misleading) relationships.
  - **Attenuation bias**: This occurs when X variables are measured with error and leads to underestimation of the regression coefficients.
- All (X, Y) observations are independent and identically distributed (i.i.d.).
- Variance of X is positive (otherwise estimation of β would not be possible).
- Variance of the errors is constant (i.e., homoskedasticity).
- It is unlikely that large outliers will be observed in the data. OLS estimates are sensitive to outliers, and large outliers have the potential to create misleading regression results.

Collectively, these assumptions ensure that the regression estimators are unbiased (i.e., E($\hat{\alpha}$) = α and E($\hat{\beta}$) = β). Secondly, they ensure that the estimators are normally distributed and, as a result, allowed for hypothesis testing (discussed later).

Properties of OLS Estimators

LO 18.d: Characterize the properties of OLS estimators and their sampling distributions.

Because OLS estimators are derived from random samples, these estimators are also random variables because they vary from one sample to the next. Therefore, OLS estimators will have their own probability distributions (i.e., sampling distributions). These sampling distributions allow us to estimate population parameters, such as the population mean, the population regression intercept term, and the population regression slope coefficient.

Drawing multiple samples from a population will produce multiple sample means. The distribution of these sample means is referred to as the *sampling distribution of the sample mean*. The mean of this sampling distribution is used as an estimator of the population mean and is said to be an **unbiased estimator** of the population mean. An unbiased estimator is one for which the expected value of the estimator is equal to the parameter you are trying to estimate.

Given the **central limit theorem (CLT)**, for large sample sizes, it is reasonable to assume that the sampling distribution will approach the normal distribution. This means that the estimator is also a **consistent estimator**. A consistent estimator is one for which the accuracy of the parameter estimate increases as the sample size increases.

Like the sampling distribution of the sample mean, OLS estimators for the population intercept term and slope coefficient also have sampling distributions. The sampling distributions of OLS estimators, α and β, are unbiased and consistent estimators of respective population parameters. Being able to assume that α and β are normally distributed is a key property in allowing us to make statistical inferences about population coefficients.

The variance of the slope (β) increases with variance of the error and decreases with the variance of the explanatory variable. This makes sense because the variance of the slope indicates the reliability of the sample estimate of the coefficient, and the higher the variance of the error, the lower the reliability of the coefficient estimate. Higher variance of the explanatory (X) variable(s) indicates that there is sufficient diversity in observations (i.e., the sample is representative of the population) and, hence, lower variability (and higher confidence) of the slope estimate.

MODULE QUIZ 18.2

1. Ordinary least squares (OLS) refers to the process that:
   A. maximizes the number of independent variables.
   B. minimizes the number of independent variables.
   C. produces sample regression coefficients.

D. minimizes the sum of the squared error terms.

2. What is the most appropriate interpretation of a slope coefficient estimate equal to 10.0?
   A. The predicted value of the dependent variable when the independent variable is zero is 10.0.
   B. The predicted value of the independent variable when the dependent variable is zero is 0.1.
   C. For every one unit change in the independent variable, the model predicts that the dependent variable will change by 10 units.
   D. For every one unit change in the independent variable, the model predicts that the dependent variable will change by 0.1 units.

3. The reliability of the estimate of the slope coefficient in a regression model is most likely:
   A. positively affected by the variance of the residuals and negatively affected by the variance of the independent variables.
   B. negatively affected by the variance of the residuals and negatively affected by the variance of the independent variables.
   C. positively affected by the variance of the residuals and positively affected by the variance of the independent variables.
   D. negatively affected by the variance of the residuals and positively affected by the variance of the independent variables.

4. The mean inflation ($\overline{Y}$) over the past 108 months is 0.01. Mean unemployment during that same time period (X) is 0.044. The variance-covariance matrix for these variables is as follows:

$$\begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix} = \begin{pmatrix} 2.54 & 45.76 \\ 45.76 & 16.84 \end{pmatrix}$$

   What is the estimated slope coefficient and intercept, respectively?
   A. 2.72 and −0.11.
   B. 1.89 and 0.01.
   C. 3.44 and −0.52.
   D. 1.44 and 1.23.

5. A researcher estimates that the value of the slope coefficient in a single explanatory variable linear regression model is equal to zero. Which one of the following is most appropriate interpretation of this result?
   A. The mean of the Y variable is zero.
   B. The intercept of the regression is zero.
   C. The relation between X and Y is not linear.
   D. The coefficient of determination (R^2) of the model is zero.

MODULE 18.3: HYPOTHESIS TESTING

LO 18.e: Construct, apply, and interpret hypothesis tests and confidence intervals for a single regression coefficient in a regression.

LO 18.f: Explain the steps needed to perform a hypothesis test in a linear regression.

LO 18.g: Describe the relationship among a *t*-statistic, its *p*-value, and a confidence interval.

The steps in the hypothesis testing procedure for regression coefficients are as follows:

1. Specify the hypothesis to be tested.
2. Calculate the test statistic.
3. Reject or fail to reject the null hypothesis after comparing the test statistic to its critical value.

Given that the OLS regression assumptions discussed previously are valid, the estimated slope coefficient, β, will be normally distributed with a standard deviation known as the **standard error of the regression** coefficient (S_b). We can then conduct hypothesis testing using sample value of the coefficient and its standard error.

Suppose we want to test the hypothesis that the value of the slope coefficient is equal to β_0.

$H_0: \beta = \beta_0$ versus $H_A: \beta \neq \beta_0$

For this example, we use the following *t*-statistic:

$$t = \frac{\beta - \beta_0}{S_b}$$

If the absolute value of the test statistic exceeds the critical *t*-value (from the *t*-table, n − 2 degrees of freedom), we would reject the null hypothesis.

EXAMPLE: Hypothesis testing of slope coefficient

A regression model estimated using 46 observations has $\beta = 0.76$ and $S_b = 0.33$. **Determine** if the slope coefficient is statistically different from zero at 5% level of significance. The critical *t*-value for a sample size of 46 and 5% level of significance is 2.02.

Answer:

$$t = \frac{\beta - \beta_0}{S_b} = \frac{0.76 - 0}{0.33} = 2.30$$

Critical *t*-value = 2.02 (given).

Because 2.30 > 2.02, we reject the null hypothesis and conclude the alternate hypothesis (that the slope coefficient is not equal to zero).

Confidence Intervals

The **confidence interval of the slope coefficient** = $\beta \pm (t_c \times S_b)$.

Where t_c is the critical *t*-value for a given level of significance and degrees of freedom (n – 2).

In the previous example, $\beta = 0.76$, $S_b = 0.33$, and $t_c = 2.02$. Thus, the confidence interval of slope coefficient = 0.76 ± (2.02 × 0.33) = 0.0934 < slope coefficient < 1.43.

Notice that zero does not fall in the confidence interval, which should always be the case if we correctly rejected the null hypothesis of β = 0 in our hypothesis test.

Similarly, we can also test the hypothesis where H_0: β = 0.20 versus H_A: β ≠ 0.20 and fail to reject the null hypothesis because 0.20 does fall within the confidence interval.

In other words, if the hypothesized value of the slope coefficient falls outside of the confidence interval, we can reject the null. If it falls inside the confidence interval, we fail to reject the null hypothesis.

The *p*-Value

The ***p*-value** is the smallest level of significance for which the null hypothesis can be rejected. An alternative method of doing hypothesis testing of regression coefficients is to compare the *p*-value to the significance level:

- If the *p*-value is less than the significance level, the null hypothesis can be rejected.
- If the *p*-value is greater than the significance level, the null hypothesis cannot be rejected.

In general, regression outputs will provide the *p*-value for the standard hypothesis (H_0: β = 0 versus H_A: β ≠ 0).

Consider again the example where β = 0.76, S_b = 0.33, and the level of significance is 5%. The regression output provides a *p*-value = 0.026. Because the *p*-value < level of significance, we reject the null hypothesis that β = 0, which is the same result as the one we got when performing the *t*-test.

MODULE QUIZ 18.3

Use the following information to answer questions 1–3.

Bob Shepperd is trying to forecast 10-year T-bond yield. Shepperd tries a variety of explanatory variables in several iterations of a single-variable model. Partial results are provided below (note that these represent three separate one-variable regressions):

Explanatory Variable	Coefficient	Standard Error	*p*-Value
Inflation	1.08	0.67	0.11
Unemployment rate	−0.48	0.12	< 0.001
GDP growth rate	1.33	0.45	0.005

The critical *t*-value at 5% level of significance is equal to 2.02.

1. For the regression model involving inflation as the explanatory variable, the confidence interval for the slope coefficient is closest to:
   A. −0.27 to 2.43.
   B. 0.26 to 2.43.
   C. −2.27 to 2.43.
   D. 0.22 to 1.88.
2. For the regression model involving unemployment rate as the explanatory variable, what are the results of a hypothesis test that the slope coefficient is equal to 0.20 (vs. not equal to 0.20) at 5% level of significance?
   A. The coefficient is not significantly different from 0.20 because the *p*-value is < 0.001.

B. The coefficient is significantly different from 0.20 because the t-value is 2.33, which is greater than the critical t-value of 2.02.
C. The coefficient is significantly different from 0.20 because the t-value is −5.67.
D. The coefficient is not significantly different from 0.20 because the t-value is −2.33.

3. For the regression model involving GDP growth rate as the explanatory variable, at a 5% level of significance, which of the following statements about the slope coefficient is least accurate?
   A. The coefficient is significantly different from 0 because the p-value is 0.005.
   B. The coefficient is significantly different from 0 because the 95% confidence interval does not include the value of 0.
   C. The coefficient is significantly different from 0 because the t-value is 2.27.
   D. The coefficient is not significantly different from 1 because t-value is 0.73.

KEY CONCEPTS

LO 18.a

Regression analysis attempts to measure the relationship between a dependent variable and one or more independent variables.

To use linear regression, the following three conditions need to be satisfied:

1. The relationship between Y and X should be linear.
2. The variance of the error term is independent of the observed data.
3. All X variables should be observable.

LO 18.b

The intercept term, α, is the value of the dependent variable when the independent variables are all equal to zero. The slope coefficient, β, is the estimated change in the dependent variable for a one-unit change in that independent variable.

$$\beta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

$$\alpha = \overline{Y} - \beta\overline{X}$$

LO 18.c

Assumptions made with linear regression include the following:

- The expected value of the error term, conditional on the independent variable, is zero.
- All (X, Y) observations are independent and identically distributed (i.i.d.).
- It is unlikely that large outliers will be observed in the data.
- The variance of X is strictly > 0.
- The variance of the errors is constant (i.e., homoskedasticity).

LO 18.d

The OLS estimators, α and β, are unbiased and consistent estimators of respective population parameters and their sampling distribution is approximately normal.

LO 18.e

To conduct tests of hypothesis for the form such as H_0: $\beta = \beta_0$ versus H_A: $\beta \neq \beta_0$

We use the following test statistic: $\dfrac{\beta - \beta_0}{S_b}$

If the absolute value of *t* exceeds the critical *t*-value (from the *t*-table, n – 2 degrees of freedom), we would reject the null hypothesis.

The confidence interval of the slope coefficient = $\beta \pm (t_c \times S_b)$

LO 18.f

Steps in hypothesis testing for linear regression:

1. Specify the hypothesis.
2. Calculate the test statistic.
3. Reject or fail to reject the null hypothesis after comparing the test statistic to its critical value.

LO 18.g

The confidence interval for the slope coefficient and *t*-test for the hypothesized value of the slope coefficient are related; if the hypothesized value falls within the confidence interval for the slope coefficient, we fail to reject the null hypothesis. If the *p*-value is less than the significance level, the null hypothesis can be rejected, otherwise we fail to reject the null.

LO 18.h

The R^2 represents the proportion of variation in the dependent variable that is explained by the independent variable(s).

$R^2 = r^2_{X,Y}$

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 18.1

1. **B** The regression equation can be written as: $E(Y) = \alpha + \beta \times X$. If $X = 0$, then $Y = \alpha$ (i.e., the intercept coefficient). (LO 18.a)

2. **A** The error term represents effects from independent variables not included in the model. It could be explained by additional independent variables. (LO 18.a)

3. **B** Linear regression refers to a regression that is linear in the coefficients/parameters; it may or may not be linear in the variables, which can enter a linear regression after appropriate transformation. (LO 18.a)

Module Quiz 18.2

1. **D** OLS is a process that minimizes the sum of squared residuals to produce estimates of the population parameters known as sample regression coefficients. (LO 18.b)

2. **C** The slope coefficient is best interpreted as the predicted change in the dependent variable for a one-unit change in the independent variable. If the slope coefficient

estimate is 10.0 and the independent variable changes by one unit, the dependent variable will change by 10 units. The intercept term is best interpreted as the value of the dependent variable when the independent variable is equal to zero. (LO 18.b)

3. **D** The reliability of the slope coefficient is inversely related to its variance and the variance of the slope coefficient (β) increases with variance of the error term and decreases with the variance of the explanatory variable. (LO 18.d)

4. **A** $$\beta = \frac{\text{Cov}(X,Y)}{\text{Var } X} = \frac{\sigma_{X,Y}}{\sigma_X^2} = \frac{45.76}{16.84} = 2.72$$

$\alpha = \overline{Y} - \beta\overline{X} = 0.01 - 2.72 \times 0.044 = -0.11$. (LO 18.b)

5. **D** When the slope coefficient is 0, variation in Y is unrelated to variation in X and correlation $r_{X,Y} = 0$. Therefore, $R^2 = r^2_{X,Y} = 0$.

Alternatively, recall that $\beta = \frac{\text{Cov}(X,Y)}{\text{Var } X}$. If $\beta = 0$, $\text{Cov}(X,Y) = 0$ and therefore:

$$r_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{0}{\sigma_X\sigma_Y} = 0$$

(LO 18.h)

Module Quiz 18.3

1. **A** The confidence interval of the slope coefficient = $\beta \pm (t_c \times S_b) = 1.08 \pm (2.02 \times 0.67)$ or −0.27 to 2.43. Notice that 0 falls within this interval and, hence, the coefficient is not significantly different from 0 at 5% level of significance. The *p*-value of 0.11 (> 0.05) also gives the same conclusion. (LO 18.e)

2. **C** The *p*-value provided is for hypothesized value of the slope coefficient being equal to 0. The hypothesized coefficient value is 0.20.

$$t = \frac{\beta - \beta_0}{S_b} = \frac{-0.48 - 0.20}{0.12} = \frac{-0.68}{0.12} = -5.67$$

(LO 18.g)

3. **C** When the *p*-value is less than the level of significance, the slope coefficient is significantly different from 0. For the test of hypothesis about coefficient value significantly different from 0:

$$t = \frac{\beta - \beta_0}{S_b} = \frac{1.33 - 0}{0.45} = 2.96$$

The confidence interval of the slope coefficient = $\beta \pm (t_c \times S_b) = 1.33 \pm (2.02 \times 0.45)$ or 0.42 to 2.34. 0 is not in this confidence interval.

For hypothesis test of coefficient is equal to 1:

$$t = \frac{\beta - \beta_0}{S_b} = \frac{1.33 - 1}{0.45} = 0.73$$

(LO 18.g)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 8.

READING 19

REGRESSION WITH MULTIPLE EXPLANATORY VARIABLES

Study Session 6

EXAM FOCUS

In this reading, we generalize the regression model to include multiple explanatory variables. For the exam, be able to evaluate and calculate goodness-of-fit measures such as R^2 and adjusted R^2 as well as hypothesis testing related to these concepts. Hypothesis testing of individual slope coefficients in a multiple regression model as well as confidence intervals of those coefficients is also important testable material.

MODULE 19.1: MULTIPLE REGRESSION

LO 19.a: Distinguish between the relative assumptions of single and multiple regression.

We extend our **regression function** in this reading to include multiple explanatory variables (which is most commonly used in practice). The general form of this **multiple regression** model is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where:
β_j = the regression or slope coefficients; sensitivity of Y to changes in X_j controlling for all other Xs
α = value of Y when all Xs = 0
ε = random error or shock; unexplained (by X) component of Y (This error may be reduced by using more independent variables or by using different, more appropriate independent variables.)

Recall the assumptions of single regression model (modified for multiple Xs):

1. The expected value of the error term, conditional on the independent variables, is zero: $[E(\varepsilon_i \mid X_{i's}) = 0]$.
2. All (Xs and Y) observations are i.i.d.
3. The variance of X is positive (otherwise estimation of β would not be possible).
4. The variance of the errors is constant (i.e., homoskedasticity).
5. There are no outliers observed in the data.

An additional sixth assumption is needed for multiple regression:

6. X variables are not perfectly correlated (i.e., they are not perfectly linearly dependent). In other words, each X variable in the model should have some variation that is not fully explained by the other X variables.

LO 19.b: Interpret regression coefficients in a multiple regression.

For a multiple regression, the interpretation of the slope coefficient is that it captures the change in the dependent variable for a one-unit change in the independent variable, holding the other independent variables constant. As a result, the slope coefficients in a multiple regression are sometimes called **partial slope coefficients**.

The **ordinary least squares (OLS)** estimation process for multiple regression differs from single regression. In a stepwise fashion, first, the individual explanatory variables are regressed against other explanatory variables and the residuals from these models become explanatory variables in the regression using the original independent variable.

Consider a simple, two-independent-variable model:

$$Y_i = \alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \varepsilon_i$$

In step 1, we estimate the residuals in the following model using OLS estimation techniques discussed previously (the estimated coefficients *a* and *b* are not actually used) for a single regression:

$$X_{1i} = a + b \times X_{2i} + \varphi_i$$

In step 2, we do the same, but this time estimate the residuals in the model:

$$Y_i = c + d \times X_{2i} + \lambda_i$$

Finally, the residuals from step 2 are regressed against the residuals from step 1 to estimate the slope coefficient β_1:

$$\lambda_i = \beta_1 \times \varphi_i + \varepsilon_i$$

This stepwise estimation process ensures that the slope coefficient (β_1) is calculated after controlling for the variation in the other independent variable (X_2). By reversing the process, we can similarly estimate β_2 after controlling for X_1.

Interpreting Multiple Regression Results

Now let's discuss the interpretation of the multiple regression slope coefficients in more detail. Suppose we run a regression of the dependent variable *Y* on a single independent variable X_1 and get the following result:

$$Y = 2.0 + 4.5X_1$$

The appropriate interpretation of the estimated slope coefficient is that if X_1 increases by 1 unit, we would expect *Y* to increase by 4.5 units.

Now suppose we add a second independent variable X_2 to the regression and get the following result:

$$Y = 1.0 + 2.5X_1 + 6.0X_2$$

Notice that the estimated slope coefficient for X_1 changed from 4.5 to 2.5 when we added X_2 to the regression. We would expect this to happen most of the time when a second variable is added to the regression, unless X_2 is uncorrelated with X_1, because if X_1 increases by 1 unit, then we would expect X_2 to change as well. The multiple regression equation captures this relationship between X_1 and X_2 when predicting Y.

Now the interpretation of the estimated slope coefficient for X_1 is that if X_1 increases by 1 unit, we would expect Y to increase by 2.5 units, holding X_2 constant.

As usual, the intercept (1.0) is interpreted as the expected value of Y when all Xs are equal to zero.

EXAMPLE: Regression model with multiple explanatory variables

A researcher estimated the following three-factor model to explain the return on different portfolios:

$$R_{P,i} = 1.70 + 1.03\ R_{m,i} - 0.23R_{z,i} + 0.32\ R_{v,i}$$

(Note: Returns are expressed in percentage form.)

Calculate the following:

1. The return on a portfolio when $R_m = 8\%$, $R_z = 2\%$ and $R_v = 3\%$
2. The impact on portfolio return if R_z declines by 1%
3. The expected return on the portfolio when $R_m = R_z = R_v = 0$

Answer:

1. $E(R_P) = 1.70 + (1.03 \times 8) - (0.23 \times 2) + (0.32 \times 3) = 10.44\%$
2. Change in portfolio return = $\Delta R_P = -\ (0.23 \times -1) = +0.23\%$
3. $E(R_P)$ when all factors = 0 would be the intercept = 1.70%

MODULE QUIZ 19.1

Use the following information to answer Questions 1 and 2.

Multiple regression was used to explain stock returns using the following variables:

Dependent variable:

RET = annual stock returns (%)

Independent variables:

MKT = market capitalization = market capitalization / $1.0 million.

IND = industry quartile ranking (IND = 4 is the highest ranking)

FORT = Fortune 500 firm, where {FORT = 1 if the stock is that of a Fortune 500 firm, FORT = 0 if not a Fortune 500 stock}

The regression results are presented in the following table.

	Coefficient	Standard Error	*t*-Statistic	*p*-Value
Intercept	0.5220	1.2100	0.430	0.681
Market capitalization	0.0460	0.0150	3.090	0.021
Industry ranking	0.7102	0.2725	2.610	0.040
Fortune 500	0.9000	0.5281	1.700	0.139

1. Based on the results in the table, which of the following most accurately represents the regression equation?
   A. 0.43 + 3.09(MKT) + 2.61(IND) + 1.70(FORT).
   B. 0.681 + 0.021(MKT) + 0.04(IND) + 0.139(FORT).
   C. 0.522 + 0.0460(MKT) + 0.7102(IND) + 0.9(FORT).
   D. 1.21 + 0.015(MKT) + 0.2725(IND) + 0.5281(FORT).
2. The expected amount of the stock return attributable to it being a Fortune 500 stock is closest to:
   A. 0.522.
   B. 0.046.
   C. 0.710.
   D. 0.900.
3. Which of the following is not an assumption of single regression?
   A. There are no outliers in the data.
   B. The variance of the independent variables is greater than zero.
   C. Independent variables are not perfectly correlated.
   D. Residual variance are homoskedastic.

MODULE 19.2: MEASURES OF FIT IN LINEAR REGRESSION

LO 19.c: Interpret goodness-of-fit measures for single and multiple regressions, including R^2 and adjusted R^2.

LO 19.e: Calculate the regression R^2 using the three components of the decomposed variation of the dependent variable data: the explained sum of squares, the total sum of squares and the residual sum of squares.

The **standard error of the regression (SER)** measures the uncertainty about the accuracy of the predicted values of the dependent variable. Graphically, the relationship

is stronger when the actual x,y data points lie closer to the regression line (i.e., the *errors* are smaller).

Recall that OLS estimation minimizes the sum of the squared differences between the predicted value and actual value for each observation. Also, recall that the regression model seeks to explain the variation in Y:

$$\Sigma(Y_i - \overline{Y})^2$$

We can write the deviation from mean for Y as:

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

Therefore,

$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma(Y_i - \hat{Y})^2$$

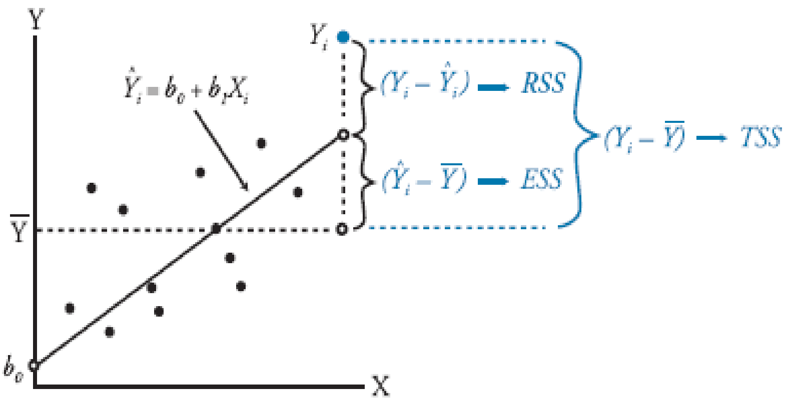
or

TSS = ESS + RSS

where:
TSS = total sum of squares (i.e., total variation in Y)
ESS = explained sum of squares (i.e., the variation in Y explained by the regression model)
RSS = residual sum of squares (i.e., the unexplained variation in Y)

Figure 19.1 illustrates how the total variation in the dependent variable (TSS) is composed of RSS and ESS.

Figure 19.1: Components of the Total Variation



Coefficient of Determination

Dividing both sides by TSS, we see that 1 = (ESS/TSS) + (RSS/TSS)

The first term on the right side captures the proportion of variation in Y that is explained. This proportion is the **coefficient of determination** (R^2) of a multiple regression and is a goodness-of-fit measure.

R^2 = ESS/TSS = % of variation explained by the regression model

Recall that for a single regression, $R^2 = r^2_{X,Y}$. For a multiple regression, $R^2 = r^2(_{Y,\hat{Y}})$.

For a multiple regression, the coefficient of determination R^2 is the square of the correlation between Y and predicted value of Y. While it is a goodness-of-fit measure, R^2 by itself may not be a reliable measure of the explanatory power of the multiple regression model due to three reasons. First, R^2 almost always increases as independent variables are added to the model, even if the marginal contribution of the new variables is not statistically significant. Consequently, a relatively high R^2 may reflect the impact of a large set of independent variables rather than how well the set explains the dependent variable. This problem is often referred to as overestimating the regression.

Adjusted R^2

To overcome the problem of overestimating the impact of additional variables on the explanatory power of a regression model, many researchers recommend adjusting R^2 for the number of independent variables. The **adjusted R^2** value is expressed as:

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:
n = number of observations
k = number of independent variables

Note that R_a^2 will be less than or equal to R^2. So, while adding a new independent variable to the model will increase R^2, it may either increase or decrease the R_a^2. If the new variable has only a small effect on R^2, the value of R_a^2 may decrease. In addition, R_a^2 may be less than zero if the R^2 is low enough.

Second, R^2 is not comparable across models with different dependent (i.e., Y) variables. Finally, there are no clear predefined values of R^2 that indicate whether the model is good or not. For some noisy variables (e.g., currency values), even models with a low R^2 may provide valuable insight.

EXAMPLE: Calculating R^2 and adjusted R^2

An analyst runs a regression of monthly value-stock returns on 5 independent variables over 60 months. The total sum of squares for the regression is 460, and the residual sum of squares is 170. **Calculate** the R^2 and adjusted R^2.

Answer:

$$R^2 = \frac{460 - 170}{460} = 0.630 = 63.0\%$$

$$R_a^2 = 1 - \left[\left(\frac{60-1}{60-5-1}\right) \times (1 - 0.63)\right] = 0.596 = 59.6\%$$

The R^2 of 63% suggests that the five independent variables together explain 63% of the variation in monthly value-stock returns.

EXAMPLE: Interpreting adjusted R^2

Suppose the analyst now adds four more independent variables to the previous regression, and the R^2 increases to 65.0%. **Identify** which model the analyst would most likely prefer.

Answer:

With nine independent variables, even though the R^2 has increased from 63% to 65%, the adjusted R^2 has decreased from 59.6% to 58.7%:

$$R_a^2 = 1 - \left[\left(\frac{60-1}{60-9-1}\right) \times (1-0.65)\right] = 0.587 = 58.7\%$$

The analyst would prefer the first model because the adjusted R^2 is higher and the model has five independent variables as opposed to nine.

Joint Hypothesis Tests and Confidence Intervals

LO 19.d: Construct, apply, and interpret joint hypothesis tests and confidence intervals for multiple coefficients in a regression.

As with single regression, the magnitude of the coefficients in a multiple regression tells us nothing about the importance of the independent variable in explaining the dependent variable. Thus, we must conduct hypothesis testing on the estimated slope coefficients to determine if the independent variables make a significant contribution to explaining the variation in the dependent variable.

The *t*-statistic used to test the significance of the individual coefficients in a multiple regression is calculated using the same formula that is used with single regression:

$$t = \frac{b_j - B_j}{s_{b_j}} = \frac{\text{estimated regression coefficient} - \text{hypothesized value}}{\text{coefficient standard error of } b_j}$$

For a multiple regression, the *t*-statistic has $(n - k - 1)$ **degrees of freedom.**

Determining statistical significance

The most common hypothesis test done on the regression coefficients is to test **statistical significance**, which means testing the null hypothesis that the coefficient is zero versus the alternative that it is not:

testing statistical significance ⇒ H_0: $b_j = 0$ versus H_A: $b_j \neq 0$

EXAMPLE: **Testing the statistical significance of a regression coefficient**

Consider the hypothesis that future 10-year real earnings growth in the S&P 500 (EG10) can be explained by the trailing dividend payout ratio of the stocks in the index (PR) and the yield curve slope (YCS). **Test** the statistical significance of the independent variable PR in the real earnings growth example at the 10% significance level. Assume that the number of observations is 46 and the critical *t*-value for 10%

level of significance is 1.68. The results of the regression are produced in the following table.

Coefficient and Standard Error Estimates for Regression of EG10 on PR and YCS

	Coefficient	Standard Error
Intercept	−11.6%	1.657%
PR	0.25	0.032
YCS	0.14	0.280

Answer:

We are testing the following hypothesis:

H_0: PR = 0 versus H_A: PR ≠ 0

The *t*-statistic is:

$$t = \frac{0.25}{0.032} = 7.8$$

Therefore, because the *t*-statistic of 7.8 is greater than the upper critical *t*-value of 1.68, we can reject the null hypothesis and conclude that the PR regression coefficient is statistically significantly different from zero at the 10% significance level.

Similar to single regression, the confidence interval for a regression coefficient in multiple regression is calculated as:

$$b_j \pm \left(t_c \times s_{b_j}\right)$$

For models with multiple variables, the univariate *t*-test is not applicable when testing complex hypotheses involving the impact of more than one variable. Instead, we use the ***F*-test**.

The F-test

An *F*-test is useful to evaluate a model against other competing partial models. For example, a model with three independent variables (X_1, X_2, and X_3) can be compared against a model with only one independent variable (X_1). We are trying to see if the two additional variables (X_2 and X_3) in the full model contribute meaningfully to explain the variation in Y.

H_0: $\beta_2 = \beta_3 = 0$ versus H_A: either $\beta_2 \neq 0$ or $\beta_3 \neq 0$

The ***F*-statistic** for multiple regression coefficients, which is always a one-tailed test, is calculated as:

$$F = \frac{(RSS_P - RSS_F)/q}{RSS_F/(n - k_F - 1)} = \frac{(R_F^2 - R_P^2)/q}{(1 - R_F^2)/(n - k_F - 1)}$$

where:

RSS_F = residual sum of squares of the full model

RSS_P = residual sum of squares of the partial model

R_F^2 = coefficient of determination of the full model

R_P^2 = coefficient of determination of the partial model

q = number of restrictions imposed on the full model to arrive at the partial model

n = number of observations

k_F = number of independent variables in the full model

The calculated *F*-statistic is compared to the critical *F*-value [with *q* degrees of freedom in the numerator and $(n - k_F - 1)$ degrees of freedom in the denominator]. If the calculated *F*-stat is greater than the critical *F*-value, the full model contributes meaningfully to explaining the variation in Y.

EXAMPLE: *F*-test

A researcher is seeking to explain returns on a stock using the market returns as an explanatory variable (CAPM formulation). The researcher wants to determine whether two additional explanatory variables contribute meaningfully to variation in the stock's return. Using a sample consisting of 64 observations, the researcher found that RSS in the model with three explanatory variables is 6,650 while the RSS in the single-variable model is 7,140. **Evaluate** the model with extra variables relative to the standard CAPM formulation.

Answer:

Given, RSS_F = 6,650; RSS_P = 7,140; n = 64; k_F = 3; and q = number of variables removed = 2, the *F*-test statistic is computed as:

$$F = \frac{(7{,}140 - 6{,}650)/2}{6{,}650/(64 - 3 - 1)} = 2.21$$

Critical *F*-value at 5% level of significance (df numerator = 2, df denominator = 60) = 3.15.

Because F = 2.21 < 3.15, we can state that the full model does not contribute meaningfully to explaining variation in Y. In other words, we fail to reject the null hypothesis that $\beta_2 = \beta_3 = 0$. Note that one of the two variables removed from the full model may still be significant, but we are only concluding here that both variables are insignificant.

A more generic *F*-test is used to test the hypothesis that all variables included in the model do not contribute meaningfully in explaining the variation in Y versus at least one of the variables does contribute statistically significantly.

H_0: $\beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0$ versus H_A: at least one $\beta_j \neq 0$

In such a case, we calculate the *F*-statistic as follows:

$$F = \frac{ESS/k}{RSS/(n-k-1)}$$

The calculated *F*-statistic is then compared to critical *F*-value (with numerator degrees of freedom = k and denominator degrees of freedom = n − k − 1). If *F*-stat > critical *F*, we reject the null hypothesis.

EXAMPLE: Calculating and interpreting the *F*-statistic

An analyst runs a regression of monthly value-stock returns on five independent variables over 46 months. The total sum of squares is 460, and the residual sum of squares is 170. **Test** the null hypothesis at the 5% significance level (95% confidence) that all five of the independent variables are equal to zero.

Answer:

The null and alternative hypotheses are:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ versus H_A: at least one $\beta_j \neq 0$

ESS = TSS − RSS = 460 − 170 = 290

$$F = \frac{290/5}{170/(46-5-1)} = \frac{58}{4.25} = 13.65$$

The critical *F*-value for 5 and 40 degrees of freedom at a 5% significance level is 2.45. Therefore, we can reject the null hypothesis and conclude that at least one of the five independent variables is significantly different than zero.

MODULE QUIZ 19.2

Use the following information to answer Questions 1 and 2.

Phil Ohlmer estimates a cross sectional regression in order to predict price to earnings ratios (P/E) with fundamental variables that are related to P/E, including dividend payout ratio (DPO), growth rate (G), and beta (B). In addition, all 50 stocks in the sample come from two industries, electric utilities or biotechnology. He defines the following dummy variable:

IND = 0 if the stock is in the electric utilities industry
or
= 1 if the stock is in the biotechnology industry

The results of his regression are shown in the following table.

Variable	Coefficient	*t*-Statistic
Intercept	6.75	3.89*
IND	8.00	4.50*
DPO	4.00	1.86
G	12.35	2.43*
B	−0.50	1.46

*Significant at the 5% level

1. Based on these results, it would be most appropriate to conclude that:
   A. biotechnology industry P/Es are statistically significantly larger than electric utilities industry P/Es.
   B. electric utilities P/Es are statistically significantly larger than biotechnology industry P/Es, holding DPO, G, and B constant.
   C. biotechnology industry P/Es are statistically significantly larger than electric utilities industry P/Es, holding DPO, G, and B constant.
   D. the dummy variable does not display statistical significance.
2. Ohlmer is valuing a biotechnology stock with a dividend payout ratio of 0.00, a beta of 1.50, and an expected earnings growth rate of 0.14. The predicted P/E on the basis of the values of the explanatory variables for the company is closest to:
   A. 7.7.
   B. 15.7.
   C. 17.2.
   D. 11.3.
3. When interpreting the R^2 and adjusted R^2 measures for a multiple regression, which of the following statements incorrectly reflects a pitfall that could lead to invalid conclusions?
   A. The R^2 measure does not provide evidence that the most or least appropriate independent variables have been selected.
   B. If the R^2 is high, we have to assume that we have found all relevant independent variables.
   C. If adding an additional independent variable to the regression improves the R^2, this variable is not necessarily statistically significant.
   D. The R^2 measure may be spurious, meaning that the independent variables may show a high R^2; however, they are not the exact cause of the movement in the dependent variable.

KEY CONCEPTS

LO 19.a

In addition to the assumptions of single regression, multiple regression requires that the X variables are not perfectly correlated (i.e., they are not perfectly linearly dependent). So, each X variable should have some variation that is not fully explained by the other X variables.

LO 19.b

For a multiple regression, the interpretation of the slope coefficient captures the change in dependent variable for one unit change in independent variable, holding the other independent variables constant.

LO 19.c and 19.e

The coefficient of determination (R^2) of a multiple regression is a goodness-of-fit measure.

R^2 = ESS/TSS = % of variation explained by the regression model
where:
TSS = total sum of squared (i.e., total variation in Y)
ESS = explained sum of squared (i.e., the variation in Y explained by the regression model)

Because R^2 almost always increases as independent variables are added to the model, to overcome the problem of overestimating the impact of additional variables we calculate adjusted R^2 as:

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:
n = number of observations
k = number of independent variables

LO 19.d

The *t*-statistic used to test the significance of the individual coefficients in a multiple regression is calculated using the same formula that is used with simple linear regression:

$$t = \frac{b_j - B_j}{s_{b_j}} = \frac{\text{estimated regression coefficient} - \text{hypothesized value}}{\text{coefficient standard error of } b_j}$$

This *t*-statistic has n – k – 1 degrees of freedom.

Similar to single regression, the confidence interval for a regression coefficient in multiple regression is calculated as:

$$b_j \pm \left(t_c \times s_{b_j}\right)$$

An *F*-test is useful to evaluate a model against other competing partial models.

$$F = \frac{(RSS_P - RSS_F)/q}{RSS_F/(n - k_F - 1)} = \frac{(R_F^2 - R_P^2)/q}{(1 - R_F^2)/(n - k_F - 1)}$$

Where *F* and *P* denote full and partial models, respectively.

A more generic *F*-test for the hypothesis: H_0: $\beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0$ versus H_A: at least one $\beta_j \neq 0$ can be conducted using the following equation:

$$F = \frac{ESS/k}{RSS/(n-k-1)}$$

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 19.1

1. **C** The coefficients column contains the regression parameters. (LO 19.b)
2. **D** The regression equation is 0.522 + 0.0460(MKT) + 0.7102(IND) + 0.9(FORT). The coefficient on FORT is the amount of the return attributable to the stock of a Fortune 500 firm. (LO 19.b)
3. **C** This is an assumption for multiple regression and not for single regression. (LO 19.a)

Module Quiz 19.2

1. **C** The *t*-statistic tests the null that industry P/Es are equal. The dummy variable is significant and positive, and the dummy variable is defined as being equal to one for biotechnology stocks, which means that biotechnology P/Es are statistically significantly larger than electric utility P/Es. Remember, however, this is only accurate if we hold the other independent variables in the model constant. (LO 19.d)
2. **B** Note that IND = 1 because the stock is in the biotech industry. Predicted P/E = 6.75 + (8.00 × 1) + (4.00 × 0.00) + (12.35 × 0.14) − (0.50 × 1.5) = 15.7. (LO 19.b)
3. **B** If the R^2 is high, we cannot assume that we have found all relevant independent variables. Omitted variables may still exist, which would improve the regression results further. (LO 19.c)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 9.

READING 20

REGRESSION DIAGNOSTICS

Study Session 6

EXAM FOCUS

This reading focuses on model specification issues and the determination of whether the assumptions underlying multiple regression are violated. For the exam, be able to explain the effects of heteroskedasticity and multicollinearity on a regression. Also, understand the bias-variance tradeoff and the consequences of including an irrelevant explanatory variable versus excluding a relevant explanatory variable.

MODULE 20.1: HETEROSKEDASTICITY AND MULTICOLLINEARITY

LO 20.a: Explain how to test whether a regression is affected by heteroskedasticity.

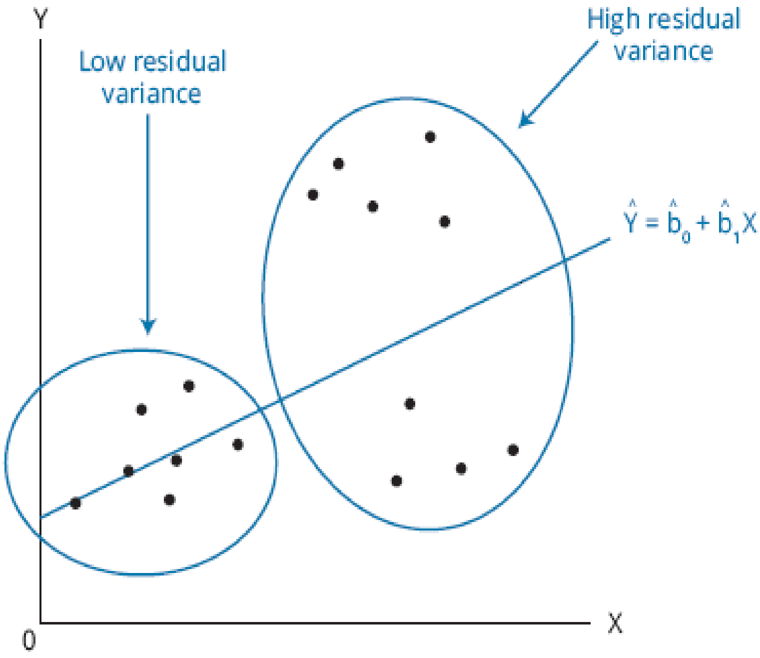
LO 20.b: Describe approaches to using heteroskedastic data.

If the variance of the residuals is constant across all observations in the sample, the regression is said to be **homoskedastic**. When the opposite is true, the regression exhibits **heteroskedasticity**, which occurs when the variance of the residuals is not the same across all observations in the sample. This happens when there are subsamples that are more spread out than the rest of the sample.

Unconditional heteroskedasticity occurs when the heteroskedasticity is not related to the level of the independent variables, which means that it doesn't systematically increase or decrease with changes in the value of the independent variable(s). While this is a violation of the equal variance assumption, it usually causes no major problems with the regression.

Conditional heteroskedasticity is heteroskedasticity that is related to the level of (i.e., conditional on) the independent variable. For example, conditional heteroskedasticity exists if the variance of the residual term increases as the value of the independent variable increases, as shown in Figure 20.1. Notice in this figure that the residual variance associated with the larger values of the independent variable, *X*, is larger than the residual variance associated with the smaller values of *X*. Conditional heteroskedasticity does create significant problems for statistical inference.

Figure 20.1: Conditional Heteroskedasticity



Effect of Heteroskedasticity on Regression Analysis

There are several effects of heteroskedasticity you need to be aware of:

- The standard errors are usually unreliable estimates.
- The coefficient estimates (i.e., the b_j) are still consistent and unbiased.
- Because of unreliable standard errors, hypothesis testing is unreliable.

Detecting Heteroskedasticity

As shown in Figure 20.1, a scatterplot of the residuals versus one of the independent variables can reveal patterns among observations. Formally, a chi-squared test statistic can be computed as follows:

1. Estimate the regression using standard ordinary least squares (OLS) procedures and estimate the residuals and square them (ε_i^2).
2. Use the squared estimated residuals in step 1 as the dependent variable in a new regression with the original explanatory variables.
3. Calculate the R^2 for the model in step 2 and use it to calculate the chi-squared test statistic:

   $$\chi^2 = nR^2$$

   The chi-squared statistic is compared to its critical value with [k × (k + 3) / 2] degrees of freedom, where k = number of independent variables.
4. If the calculated χ^2 > critical χ^2, we reject the null hypothesis of no conditional heteroskedasticity.

Correcting for Heteroskedasticity

If conditional heteroskedasticity is detected, we can conclude that the coefficients are unaffected but the standard errors are unreliable. In such a case, revised, *White standard errors* should be used in hypothesis testing instead of the standard errors from OLS estimation procedures.

PROFESSOR'S NOTE

White standard errors are heteroskedasticity-consistent standard errors. The introduction of these robust standard errors is credited to Halbert White, a well-known professor in econometrics.

LO 20.c: Characterize multicollinearity and its consequences; distinguish between multicollinearity and perfect collinearity.

Recall from the previous reading the additional assumption needed in multiple regression as opposed to a single regression: X variables are not perfectly correlated (i.e., they are not perfectly linearly dependent). When the X variables are perfectly correlated, it is called as **perfect collinearity**. This would be the case when one of the independent variables can be perfectly characterized by a linear combination of other independent variables (e.g., $X_3 = 2X_1 + 3X_2$).

Multicollinearity refers to the condition when two or more of the independent variables, or linear combinations of the independent variables, in a multiple regression are highly correlated with each other. While multicollinearity does not represent a violation of regression assumptions, its existence compromises the reliability of parameter estimates.

Effect of Multicollinearity On Regression Analysis

As a result of multicollinearity, there is a greater probability that we will incorrectly conclude that a variable is not statistically significant (e.g., a **Type II error**). Multicollinearity is likely to be present to some extent in most economic models. The issue is whether the multicollinearity has a significant effect on the regression results.

Detecting Multicollinearity

The most common way to detect multicollinearity is the situation where *t*-tests indicate that none of the individual coefficients is significantly different than zero, while the R^2 is high (and the *F*-test rejects the null hypothesis). This suggests that the variables together explain much of the variation in the dependent variable, but the individual independent variables do not. The only way this can happen is when the independent variables are highly correlated with each other, so while their common source of variation is explaining the dependent variable, the high degree of correlation also "washes out" the individual effects.

EXAMPLE: Detecting multicollinearity

Bob Watson runs a regression of mutual fund returns on average P/B, average P/E, and average market capitalization, with the following results:

Variable	Coefficient	*p*-Value
Average P/B	3.52	0.15
Average P/E	2.78	0.21
Market Cap	4.03	0.11
R^2	89.6%	

Determine whether or not multicollinearity is a problem in this regression.

Answer:

The R^2 is high, which suggests that the three variables as a group do an excellent job of explaining the variation in mutual fund returns. However, none of the independent variables individually is statistically significant to any reasonable degree, because the *p*-values are larger than 10%. This is a classic indication of multicollinearity.

Another approach to identify multicollinearity is to calculate the **variance inflation factor (VIF)** for each explanatory variable. To do that, we calculate R^2 in the model using the subject explanatory variable (X_j) as the dependent variable and the other X variables as independent variables:

$$X_j = b_0 + b_1X_1 + \dots + b_{j-1}X_{j-1} + b_{j+1}X_{j+1} + \dots + b_kX_k$$

This R^2 is then used in the VIF formula as follows:

$$VIF_j = \frac{1}{1 - R_j^2}$$

A VIF > 10 (i.e., R^2 > 90%) should be considered problematic for that variable.

Correcting Multicollinearity

The most common method to correct for multicollinearity is to omit one or more of the correlated independent variables. Unfortunately, it is not always an easy task to identify the variable(s) that are the source of the multicollinearity. There are statistical procedures that may help in this effort, like stepwise regression, which systematically remove variables from the regression until multicollinearity is minimized.

MODULE QUIZ 20.1

1. Effects of conditional heteroskedasticity include which of the following problems?
   I. The coefficient estimates in the regression model are biased.
   II. The standard errors are unreliable.

   A. I only.
   B. II only.
   C. Both I and II.
   D. Neither I nor II.

2. Der-See Hsu, researcher for Xiang Li Quant Systems, is using a multiple regression model to forecast currency values. Hsu determines that the chi-squared statistics calculated using the R^2 of the regression involving the squared residuals as dependent variable exceeds the chi-squared critical value. Which of the following is the most appropriate conclusion for Hsu to reach?
   A. Hsu should estimate the White standard errors for use in hypothesis testing.

B. OLS estimates and standard errors are consistent, unbiased, and reliable.
C. OLS coefficients are biased but standard errors are reliable.
D. A linear model is inappropriate to model the variation in the dependent variable.

3. Ben Strong recently joined Equity Partners as a junior analyst. Within a few weeks, Strong successfully modeled the movement of price for a hot stock using a multiple regression model. Beth Sinclair, Strong's supervisor, is in charge of evaluating the results of Strong's model. What is the most appropriate conclusion for Sinclair based on the variance information factor (VIF) for each of the explanatory variables included in Strong's model as shown here?

Variable	VIF
X1	2.1
X2	10.3
X3	6.9

A. Variables X1 and X2 are highly correlated and should be combined into one variable.
B. Variable X3 should be dropped from the model.
C. Variable X2 should be dropped from the model.
D. Variables X1 and X2 are not statistically significant.

4. Which of the following statements regarding multicollinearity is least accurate?
A. Multicollinearity may be present in any regression model.
B. Multicollinearity is not a violation of a regression assumption.
C. Multicollinearity makes it difficult to determine the contribution to explanation of the dependent variable of an individual explanatory variable.
D. If the *t*-statistics for the individual independent variables are insignificant, yet the *F*-statistic is significant, this indicates the presence of multicollinearity.

MODULE 20.2: MODEL SPECIFICATION

LO 20.d: Describe the consequences of excluding a relevant explanatory variable from a model and contrast those with the consequences of including an irrelevant regressor.

PROFESSOR'S NOTE

A regressor is often used as a term to describe an independent (or X) variable.

Model specification is an art requiring a thorough understanding of the underlying economic theory that explains the behavior of the dependent variable. For example, many factors may influence short-term interest rates, including inflation rate, unemployment rate, GDP growth rate, capacity utilization, and so forth. Analysts trying to model a variable need to determine the factors that should be included/excluded in their model.

While including irrelevant/extraneous variables does not pose any serious challenges, the model's adjusted R^2 declines (recall that unless a variable contributes meaningfully to explaining the variation in Y, its inclusion reduces the adjusted R^2).

Omitting relevant factors from an ordinary least squares (OLS) regression can produce misleading or biased results. **Omitted variable bias** is present when two conditions

are met: (1) the omitted variable is correlated with other independent variables in the model, and (2) the omitted variable is a determinant of the dependent variable. When relevant variables are absent from a linear regression model, the results will likely lead to incorrect conclusions, as the OLS estimators may not accurately portray the actual data.

The coefficients of the included variables that are correlated with the omitted variable will partly (depending on the correlation between them) pick up the impact of the omitted variable (leading to biased estimates of coefficients of those variables). Furthermore, the uncorrelated portion of the omitted variable's influence on the dependent variable gets captured by the error, magnifying it.

The issue of omitted variable bias occurs regardless of the size of the sample and will make OLS estimators inconsistent. The correlation between the omitted variable and the included independent variables will determine the size of the bias (i.e., a larger correlation will lead to a larger bias) and the direction of the bias (i.e., whether the correlation is positive or negative). The coefficients of the included independent variables therefore would be biased and inconsistent.

Bias-Variance Tradeoff

LO 20.e: Explain two model selection procedures and how these relate to the bias-variance tradeoff.

The holy grail of model specification is selecting the appropriate explanatory variables to include in the model. Models with too many explanatory variables (i.e., overfit models) may explain the variation in dependent variable well in-sample, but perform poorly out-of-sample. Overfit, larger models have lower bias and higher variance (i.e., estimation) errors due to inclusion of too many independent variables. Smaller, less complex models, on the other hand, have higher bias and lower variance errors (i.e., lower R^2). There are two ways to deal with this bias-variance tradeoff:

1. **General-to-specific model**: involves starting with the largest model and then successively dropping independent variables that have the smallest absolute *t*-statistic.
2. ***m*-fold cross-validation**: involves dividing the sample into *m* parts and then using (m-1) parts (known as the training set) to fit the model and the remaining part (known as the validation set) to use for out-of-sample validation. A set of candidate models are first determined and then tested using this procedure to find the optimal model—one which has the lowest *out-of-sample* error.

Residual Plots

LO 20.f: Describe the various methods of visualizing residuals and their relative strengths.

Basic **residual plots** show the residuals on the y-axis and the predicted value of the dependent variable $(\hat{y})$ on the x-axis. Ideally, the residuals should be small in magnitude, and not related to any of the explanatory variables. Alternatively, standardized residuals (i.e., the residuals divided by their standard deviation) could be plotted on the y-axis. The magnitude of the residuals would then be standardized and any residual over ±4 standard deviations would be considered problematic.

Identifying Outliers

LO 20.g: Describe methods for identifying outliers and their impact.

Recall that one of the assumptions of linear regression is that there are no outliers in the sample data. This is because the presence of outliers skews the estimated regression parameters. Outliers, when removed, induce large changes in the value of the estimated coefficients. One metric to identify an outlier is **Cook's measure**, which is computed as follows:

$$D_j = \frac{\sum_{i=1}^{n}(\hat{y}_i^{(-j)} - \hat{y}_i)^2}{kS^2}$$

where:

$\hat{y}^{(-j)}$ = predicted value of y after dropping outlier observation j
$\hat{y}_i$ = predicted value of y without dropping any observation
k = number of independent variables
S^2 = squared residuals in the model with all observations

Large values of Cook's measure (i.e., $D_j > 1$) indicate that the dropped observation was indeed an outlier.

The Best Linear Unbiased Estimator

LO 20.h: Determine the conditions under which OLS is the best linear unbiased estimator.

For OLS to generate the **best linear unbiased estimator (BLUE)**, the assumptions underlying the linear regression need to be satisfied. Specifically, the relationship between Y and X(s) should be linear and residuals should be homoskedastic (i.e., residual distribution should be identical), independent, and have an expected value of zero. The last few assumptions are summarized as $\varepsilon_i \sim N(\text{i.i.d.})$.

If there are no outliers, and the residuals have an expected value of zero, we can relax the assumption of normality for the residual distribution.

MODULE QUIZ 20.2

1. The omitted variable bias results from:
   A. exclusion of uncorrelated independent variables.

B. inclusion of uncorrelated independent variables.
C. inclusion of correlated independent variables.
D. exclusion of correlated independent variables.

2. Which of the following statements about bias-variance tradeoff is most accurate?
   A. Models with a large number of independent variables tend to have a high bias error.
   B. High variance error results when the out-of-sample R^2 of a regression is high.
   C. Models with fewer independent variables tend to have a high variance error.
   D. General-to-specific model is one approach to resolve the bias-variance tradeoff.

3. Evaluate the following statements:
   I. A high value of Cook's measure indicates the presence of an outlier.
   II. Cook's measure is inversely related to the squared residuals.

   A. Both statements are correct.
   B. Only statement I is correct.
   C. Only statement II is correct.
   D. Both statements are incorrect.

KEY CONCEPTS

LO 20.a

Conditional heteroskedasticity indicates that the variance of the residual term is conditioned on the value of the independent variable. Even though the coefficient estimates are unbiased and consistent, the estimated standard errors are unreliable in the presence of conditional heteroskedasticity. The results of any hypothesis testing are therefore unreliable.

LO 20.b

If conditional heteroskedasticity is detected, we can conclude that the coefficients are unaffected but the standard errors are unreliable. In such a case, revised, White estimated standard errors should be used in hypothesis testing instead of the standard errors from OLS procedures.

LO 20.c

When the X variables are perfectly correlated, it is called perfect collinearity. Multicollinearity refers to when two or more of the independent variables, or linear combinations of the independent variables, in a multiple regression are highly correlated with each other. As a result of multicollinearity, there is a greater probability that we will incorrectly conclude that a variable is not statistically significant (e.g., a Type II error). One of the clues for presence of multicollinearity is when there is a disconnect between *t*-tests for significance of individual slope coefficients and the *F*-test for the overall model. Alternatively, the variance inflation factor (VIF) for each explanatory variable can be calculated to indicate the presence of multicollinearity; a VIF >10 for a variable indicates the presence of multicollinearity.

LO 20.d

While including irrelevant/extraneous variables does not pose any serious challenges, the model's adjusted R^2 declines.

Omitting relevant factors from an ordinary least squares (OLS) regression can produce misleading or biased results. Omitted variable bias is present when two conditions are

met: (1) the omitted variable is correlated with other independent variables in the model, and (2) the omitted variable is a determinant of the dependent variable.

LO 20.e

Bias-variance tradeoff involves selecting between overfit models with too many variables and higher complexity (i.e., high variance, but low bias) versus models with fewer explanatory variables and lower complexity (i.e., high bias, but low variance).

LO 20.f

Two methods of plotting residuals versus predicted y-values include raw residuals and standardized residuals. The benefit of using standardized residuals is that outliers can be quickly visualized when its value exceeds ±4.

LO 20.g

Apart from residual plots, outliers can be identified via Cook's measure as follows:

$$D_j = \frac{\sum_{i=1}^{n} (\hat{y}_i^{(-j)} - \hat{y}_i)^2}{kS^2}$$

LO 20.h

OLS generates best linear unbiased estimates (BLUE) when the residual variance is constant and has an expected value of zero (even if the distribution of residuals is not normal).

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 20.1

1. **B** Effects of heteroskedasticity include the following: (1) The standard errors are usually unreliable estimates and (2) the coefficient estimates are not affected. (LO 20.a)

2. **A** Hsu's test results indicate that the null hypothesis of no conditional heteroskedasticity should be rejected. In such a case, the OLS estimates of standard errors would be unreliable and Hsu should estimate White corrected standard errors for use in hypothesis testing. Coefficient estimates would still be reliable (i.e., unbiased and consistent) (LO 20.b)

3. **C** VIF > 10 for independent variable X2 indicates that it is highly correlated with the other two independent variables in the model, indicating multicollinearity. One of the approaches to overcoming the problem of multicollinearity is to drop the highly correlated variable. (LO 20.c)

4. **A** Multicollinearity will not be present in a single regression. While perfect collinearity is a violation of a regression assumption, the presence of multicollinearity is not. Divergence between *t*-test and *F*-test is one way to detect the presence of multicollinearity. Multicollinearity makes it difficult to precisely measure the contribution of an independent variable toward explaining the variation in the dependent variable. (LO 20.c)

Module Quiz 20.2

1. **D** Omitted variable bias results from excluding a relevant independent variable that is correlated with other independent variable. (LO 20.d)
2. **D** Larger, overfit models have a low bias error (high R^2 in-sample but low R^2 out-of-sample). Smaller, parsimonious models have lower R^2 in-sample and a lower variance error. Two ways to resolve the bias-variance tradeoff are the general-to-specific model and *m*-fold cross-validation. (LO 20.e)
3. **A** Both statements are correct. A high Cook measure for an observation (> 1) indicates that it is an outlier. The squared residuals are in the denominator in the computation of the Cook measure and, hence, are inversely related to the measure. (LO 20.g)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 10.

READING 21

STATIONARY TIME SERIES

Study Session 7

EXAM FOCUS

In this reading, we learn to model the cyclical component of a time series using autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) processes. For the past values of a time series to serve as a guide for its future values, it is necessary that the time series is stationary (i.e., past patterns are expected to continue). For the exam, know the difference between an AR process and an MA process and how some series can be modeled best with a combination of the two. Finally, understand the model evaluation using residual autocorrelations.

MODULE 21.1: COVARIANCE STATIONARY

A **time series** is data collected over regular time periods (e.g., monthly S&P 500 returns, quarterly dividends paid by a company, etc.). Time series data have trends (the component that changes over time), seasonality (systematic change that occur at specific times of the year), and cyclicality (changes occurring over time cycles). For this reading, we are concerned with the third component. This cyclical component can be decomposed into *shocks* and *persistence components*. While we discuss the seasonal component briefly at the end of the reading, for the most part, we will limit ourselves to linear models to model the persistence component.

LO 21.a: Describe the requirements for a series to be covariance-stationary.

A process such as a time series must have certain properties if we want to forecast its future values based on its past values. In particular, it needs the relationships among its present and past values to remain stable over time. We refer to such a time series as being **covariance stationary**.

To be covariance stationary, a time series must exhibit the following three properties:

1. Its mean must be stable over time.
2. Its variance must be finite and stable over time.
3. Its covariance structure must be stable over time.

Covariance structure refers to the covariances among the values of a time series at its various **lags**, which are a given number of periods apart at which we can observe its values. We use the lowercase Greek letter tau, τ, to represent a lag. For example, τ = 1 refers to a one-period lag, comparing each value of a time series to its preceding value, and if τ = 4 we are comparing values four periods apart along the time series.

Autocovariance and Autocorrelation Functions

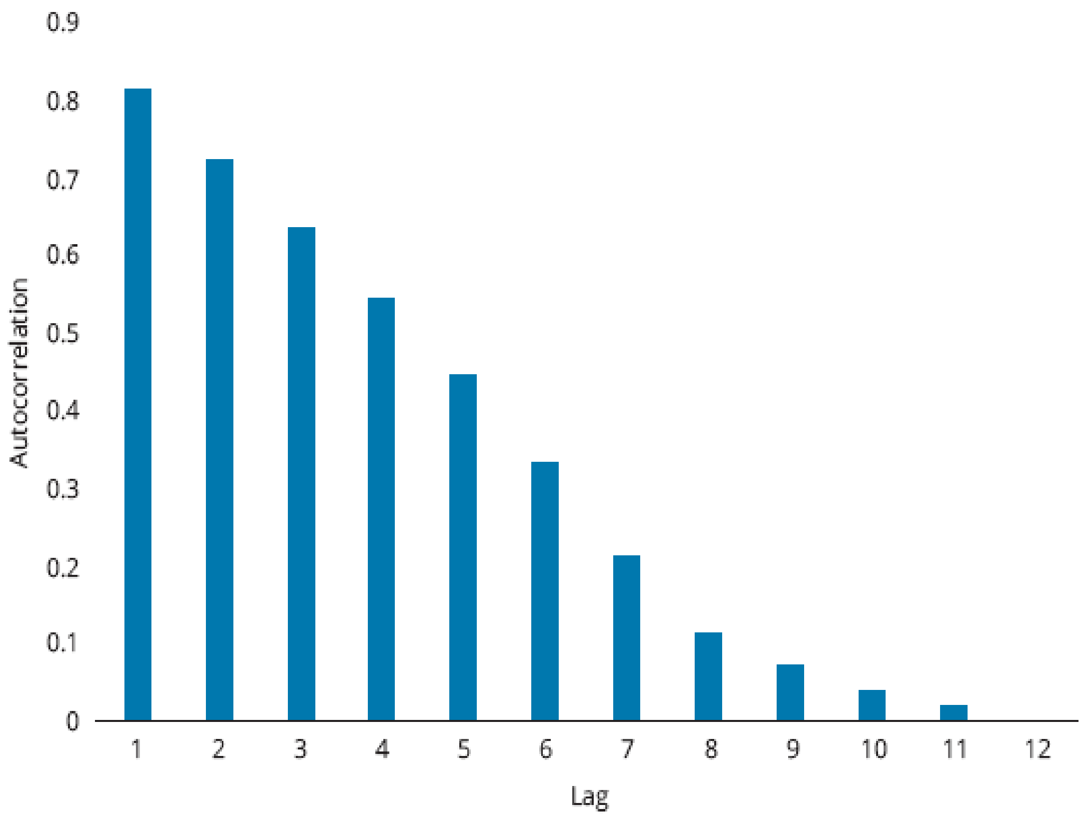
LO 21.b: Define the autocovariance function and the autocorrelation function.

The covariance between the current value of a time series and its value τ periods in the past is referred to as its **autocovariance** at lag τ. Its autocovariances for all τ make up its **autocovariance function**. If a time series is covariance stationary, its autocovariance function is stable over time. That is, its autocovariance depends on the τ we choose, but does not depend on the time over which we observe the series.

As we often do when working with covariances, we can convert them to correlations to better interpret the strength of the relationships. To convert an autocovariance function to an **autocorrelation function (ACF)**, we divide the autocovariance at each τ by the variance of the time series. This gives us an autocorrelation for each τ that will be scaled between −1 and +1.

A useful way to analyze an ACF is to display it on a graph. Figure 21.1 illustrates an example of an ACF. As can be seen in the graph, the autocorrelations approach zero as τ gets large. This is always the case for a covariance stationary series.

Figure 21.1: Autocorrelation Function



A related function is the **partial autocorrelation function**, which makes up the correlations for all lags after *controlling* for the values between the lags (think about coefficient values in a regression when including all the lags as independent variables).

PROFESSOR'S NOTE

These are *partial* in the sense that they are regressed one lag at a time. For example, if we regress a monthly time series against its year-ago values, we get a partial autocorrelation for $\tau = 12$ that does not account for any effects from other lags. We would be unlikely to get the same result for $\tau = 12$ if we ran a multiple regression that also included $\tau = 1$, $\tau = 2$, and so forth.

While autocorrelations successively decline, it is not so for partial autocorrelations; partial autocorrelations experience a steep decline. Partial autocorrelations may be large only for a few lags and those lags become prime candidates for inclusion in an autoregressive (AR) model, discussed later.

White Noise

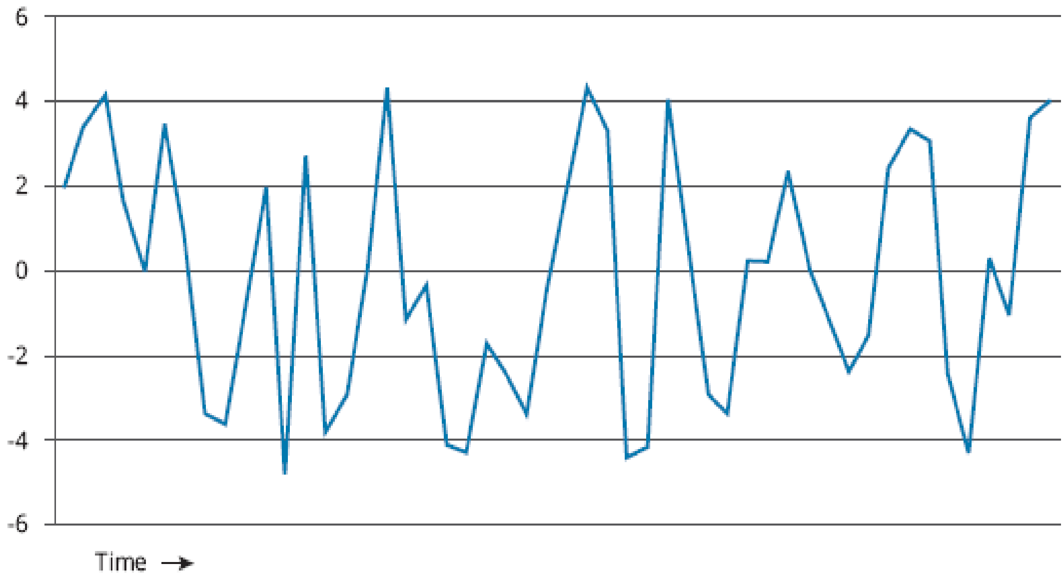
LO 21.c: Define white noise, and describe independent white noise and normal (Gaussian) white noise.

A time series might exhibit zero correlation among any of its lagged values. Such a time series is said to be **serially uncorrelated**. A special type of serially uncorrelated series is one that has a mean of zero and a constant variance. This condition is referred to as **white noise**, or zero-mean white noise, and the time series is said to follow a white noise process.

If the observations in a white noise process are independent, as well as uncorrelated, the process is referred to as **independent white noise**. If the process also follows a normal distribution, it is known as **normal white noise** or **Gaussian white noise**. Not all independent white noise processes are normally distributed, but all normal white noise processes are also independent white noise.

Graphically, a white noise process resembles Figure 21.2, with no identifiable patterns among the time periods.

Figure 21.2: White Noise Process



One important purpose of the white noise concept is to analyze a forecasting model. A model's forecast errors should follow a white noise process. If they do not, the errors themselves can be forecasted based on their past values. This implies that the model is inaccurate in a predictable way and is therefore inadequate; it needs to be revised, perhaps by adding more lags.

Earlier, we stated that a white noise process has a mean of zero and a constant variance; this refers to its *unconditional mean and variance*. A process may have a *conditional mean and variance* that are not necessarily constant. That is, the expected value of the next observation in the series might not be the mean of the time series if the next observation is conditional on one or more of its earlier values. If such a relationship exists, we can use it for forecasting the time series.

For an *independent* white noise process, we can say the next value in the series has no conditional relationship to any of its past values. Therefore, its conditional mean is the same as its unconditional mean. In this case, we cannot forecast based on past values.

Wold's theorem proposes a way to model the role of white noise and holds that a covariance stationary process can be modeled as an infinite distributed lag of a white noise process. Such a model would take the following form:

$$y_t = \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \ldots = \sum_{i=0}^{\infty} b_i\varepsilon_{t-i}$$

Where the *b* variables are constants and ε_t is a white noise process.

Because this expression can be applied to any covariance stationary series, it is known as a **general linear process**.

MODULE QUIZ 21.1

1. The conditions for a time series to exhibit covariance stationarity are least likely to include:
   A. a stable mean.
   B. a finite variance.
   C. a finite number of observations.
   D. autocovariances that do not depend on time.
2. As the number of lags or displacements becomes large, autocorrelation functions (ACFs) will approach:

A. −1.
B. 0.
C. 0.5.
D. +1.

3. Which of the following statements about white noise is most accurate?
   A. All serially uncorrelated processes are white noise.
   B. All Gaussian white noise processes are independent white noise.
   C. All independent white noise processes are Gaussian white noise.
   D. All serially correlated Gaussian processes are independent white noise.

MODULE 21.2: AUTOREGRESSIVE AND MOVING AVERAGE MODELS

Autoregressive Processes

LO 21.d: Define and describe the properties of autoregressive (AR) processes.

LO 21.g: Explain mean reversion and calculate a mean-reverting level.

LO 21.m: Describe the role of mean reversion in long-horizon forecasts.

Autoregressive models are the most widely applied time series models in finance. The **first-order autoregressive [AR(1)] process** is specified in the form of a variable regressed against itself in lagged form. This relationship can be shown in the following formula:

$$y_t = d + \Phi y_{t-1} + \varepsilon_t$$

where:
d = intercept term
y_t = the time series variable being estimated
y_{t-1} = one-period lagged observation of the variable being estimated
ε_t = current random white noise shock (mean 0)
Φ = coefficient for the lagged observation of the variable being estimated

In order for an AR(1) process to be covariance stationary, the absolute value of the coefficient on the lagged operator must be less than one (i.e., $|\Phi| < 1$). Similarly, for an AR(p) process, the sum of all coefficients should be less than 1.

The long-run (or unconditional) mean reverting level of an AR(1) series =

$$\mu = \frac{d}{1 - \Phi}$$

The long-run (or unconditional) mean reverting level of an AR(p) series =

$$\mu = \frac{d}{1 - \Phi_1 - \Phi_2 \ldots - \Phi_p}$$

This mean reverting level acts as an attractor such that the time series moves toward its mean over time.

For an AR(1) process, the variance of $y_t = \frac{\sigma_\varepsilon^2}{1 - \Phi^2}$. Similarly, we can calculate the variance of an AR(p) process by subtracting all the squared coefficients in the denominator.

For example, if we are modeling daily demand for ice cream, we would forecast our current period daily demand (y_t) as a function of a coefficient (Φ) multiplied by our lagged daily demand for ice cream (y_{t_1}) and then add a random error shock (ε_t). This process enables us to use a past observed variable to predict a current observed variable.

To estimate the autoregressive parameters, such as the coefficient (Φ), forecasters need to accurately estimate the autocovariance function of the data series:

$$y_t = \Phi^{|t|} y_0$$

The **Yule-Walker equation** is used for this purpose. When using the Yule-Walker concept to solve for the autocorrelations of an AR(1) process, we use the following relationship:

$$\rho_t = \Phi^{|t|} \text{ for } t = 0,1,2,\ldots$$

The significance of the Yule-Walker equation is that for autoregressive processes, the autocorrelation decays geometrically to zero as *t* increases.

Consider an AR(1) process that is specified using the following formula:

$$y_t = 0.65y_{t-1} + \varepsilon_t$$

The coefficient (Φ) is equal to 0.65; the first-period autocorrelation is 0.65 (i.e., 0.65^1); the second-period autocorrelation is 0.4225 (i.e., 0.65^2); and so forth for the remaining autocorrelations.

It should also be noted that if the coefficient (Φ) were to be a negative number, perhaps −0.65, then the decay would still occur, but the value would oscillate between negative and positive numbers. This is true because $-0.65^3 = -0.2746$, $-0.65^4 = 0.1785$, and $-0.65^5 = -0.1160$. You would still notice the absolute value decaying, but the actual autocorrelations would alternate between positive and negative numbers over time.

Moving Average (MA) Processes

LO 21.e: Define and describe the properties of moving average (MA) processes.

Conceptually, an MA process is a linear regression of the current values of a time series against both the current and previous unobserved white noise error terms, which are random shocks. MAs are always covariance stationary. The **first-order moving average [MA(1)] process** can be defined as:

$$y_t = \mu + \theta\varepsilon_{t-1} + \varepsilon_t$$

where:

μ = mean of the time series
ε_t = current random white noise shock (mean 0)
ε_{t-1} = one-period lagged random white noise shock
θ = coefficient for the lagged random shock

The MA(1) process is considered to be *first-order* because it only has one lagged error term (ε_{t_1}). This yields a very short-term memory because it only incorporates what happens one period ago. If we ignore the lagged error term for a moment and assume that $\varepsilon_t > 0$, then $y_t > 0$. This is equivalent to saying that a positive error term will yield a positive dependent variable (y_t). When adding back the lagged error term, we are now saying that the dependent variable is impacted by not only the current error term, but also the previous period's unobserved error term, which is amplified by a coefficient (θ). Consider an example using daily demand for ice cream (y_t) to better understand how this works:

$$y_t = 5{,}000 + 0.3\varepsilon_{t-1} + \varepsilon_t$$

In this equation, the error term is the daily change in demand. Using only the current period's error term (ε_t), if the daily change is positive, then we would estimate that daily demand for ice cream would also be positive. But, if the daily change yesterday (ε_{t_1}) was also positive, then we would expect an amplified impact on our daily demand by a factor of 0.3. If the coefficient θ is negative, the series aggressively mean reverts because the effect of the previous shock reverts in the current period.

One key feature of MA processes is called the **autocorrelation (ρ) cutoff**. We would compute the autocorrelation using the following formula:

$$\rho_1 = \frac{\theta_1}{1+\theta_1^2}\text{; where } \rho_\tau = 0 \text{ for } \tau > 1$$

Using the previous example with $\theta = 0.3$, we would compute the autocorrelation to be 0.2752 as follows:

$$0.2752 = \frac{0.3}{1+0.3^2}$$

For any value beyond the first lagged error term, the autocorrelation will be zero in an MA(1) process. This is important because it is one condition of being covariance stationary (i.e., mean = 0, variance = σ^2), which is a condition of this process being a useful estimator.

It is also important to note that this **moving average representation** has both a current random shock (ε_t) and a lagged unobservable shock (ε_{t_1}) on the independent side of this equation. This presents a problem for forecasting in the real world because it does not incorporate observable shocks.

A more general form of a moving average process, MA(q), incorporates *q* lags:

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t_1} + \ldots.. + \theta_q\varepsilon_{t_q}$$

The mean of the MA(q) process is still μ but the variance will change to:

$\sigma_y^2 = \sigma_\varepsilon^2(1 + \theta^2{}_1 + \theta^2{}_2 + ... + \theta^2{}_q)$

Lag Operators

LO 21.f: Explain how a lag operator works.

A commonly used notation for time series modeling is the **lag operator** (L). If y_t is the value of a time series at time *t*, and y_{t-1} is its value one period earlier, we can express a lag operator as:

$$y_{t-1} = Ly_t$$

There are six properties of the lag operator:

1. It shifts the time index back by one period.
2. To apply the lag operator over multiple periods, $L^m y_t = y_{t-m}$.
3. When applied to a constant, the lag operator does not change the constant.
4. Forecasting models often take the form of a **distributed lag** that assigns weights to the past values of a time series. For example, suppose we have the following model:

   $$y_t + 0.7y_{t-1} + 0.4y_{t-2} + 0.2y_{t-3}$$

   Using lag operators in this model (known as a lag polynomial), it would be expressed as:

   $$(1 + 0.7L + 0.4L^2 + 0.2L^3)y$$
5. Lag polynomials can be multiplied.
6. Assuming that the coefficients satisfy some conditions, the polynomial can be inverted.

   There are two main purposes of using a lag operator. First, an AR process is covariance stationary only if its lag polynomial is invertible. Second, this invertibility is used to select the appropriate time series model among equivalent models by applying what is known as the Box-Jenkins methodology.

MODULE QUIZ 21.2

1. Which of the following conditions is necessary for an autoregressive (AR) process to be covariance stationary?
   A. The value of the lag slope coefficients should add to 1.
   B. The value of the lag slope coefficients should all be less than 1.
   C. The absolute value of the lag slope coefficients should be less than 1.
   D. The sum of the lag slope coefficients should be less than 1.
2. Which of the following statements is a key differentiator between a moving average (MA) representation and an autoregressive (AR) process?
   A. An MA representation shows evidence of autocorrelation cutoff.
   B. An AR process shows evidence of autocorrelation cutoff.
   C. An unadjusted MA process shows evidence of gradual autocorrelation decay.
   D. An AR process is never covariance stationary.

3. Assume in an autoregressive [AR(1)] process that the coefficient for the lagged observation of the variable being estimated is equal to 0.75. According to the Yule-Walker equation, what is the second-period autocorrelation?
   A. 0.375.
   B. 0.5625.
   C. 0.75.
   D. 0.866.
4. Which of the following statements is most likely a purpose of the lag operator?
   A. A lag operator ensures that the parameter estimates are consistent.
   B. An autoregressive (AR) process is covariance stationary only if its lag polynomial is invertible.
   C. Lag polynomials can be multiplied.
   D. A lag operator ensures that the parameter estimates are unbiased.

MODULE 21.3: AUTOREGRESSIVE MOVING AVERAGE (ARMA) MODELS

LO 21.h: Define and describe the properties of autoregressive moving average (ARMA) processes.

So far, we have examined MA processes and AR processes assuming they interact independently of each other. While this may be the case, it is possible for a time series to show signs of both processes and theoretically capture a still richer relationship. For example, stock prices might show evidence of being influenced by both unobserved shocks (the MA component) and their own lagged behavior (the autoregressive component). This more complex relationship is called an **autoregressive moving average (ARMA) process** and is expressed by the following formula:

$$y_t = d + \Phi y_{t_1} + \varepsilon_t + \theta\varepsilon_{t_1}$$

where:
d = intercept term
y_t = the time series variable being estimated
Φ = coefficient for the lagged observations of the variable being estimated
y_{t_1} = one-period lagged observation of the variable being estimated
ε_t = current random white noise shock
θ = coefficient for the lagged random shocks
ε_{t_1} = one-period lagged random white noise shock

You can see that the ARMA specification merges the concepts of an AR process and an MA process. In order for the ARMA process to be covariance stationary, which is important for forecasting, we must still observe that $|\Phi| < 1$. Just as with the AR process, the autocorrelations in an ARMA process will also decay gradually for essentially the same reasons.

Consider an example regarding sales of an item (y_t) and a random shock of advertising (ε_t). We could attempt to forecast sales for this item as a function of the previous period's sales (y_{t_1}), the current level of advertising (ε_t), and the one-period lagged level of advertising (ε_{t_1}). It makes intuitive sense that sales in the current period could be

affected by both past sales and by random shocks, such as advertising. Another possible random shock for sales could be the seasonal effects of weather conditions.

PROFESSOR'S NOTE

Just as MA models can be extrapolated to the q^{th} observation and AR models can be taken out to the p^{th} observation, ARMA models can be used in the format of an ARMA(p,q) model. For example, an ARMA(3,1) model means three lagged operators in the AR portion of the formula and one lagged operator on the MA portion. This flexibility provides the highest possible set of combinations for time series forecasting of the three models discussed in this reading.

Application of AR, MA, and ARMA Processes

LO 21.i: Describe the application of AR, MA, and ARMA processes.

LO 21.l: Explain how forecasts are generated from ARMA models.

A forecaster might begin by plotting the autocorrelations for a data series and find that the autocorrelations cut off abruptly. In this case, the forecaster should consider using an MA process. If the autocorrelations instead decay gradually, he should consider using either an AR process or an ARMA process. The forecaster should especially consider these alternatives if he notices periodic spikes in the autocorrelations as they are gradually decaying. For example, if every 12^{th} autocorrelation jumps upward, this observation indicates a possible seasonality effect in the data and would heavily point toward using either an AR or ARMA model.

Another way of looking at model applications is to test various models using regression results. It is easiest to see the differences using data that follows some pattern of seasonality, such as employment data. In the real world, a moving average process would not specify a very robust model, and autocorrelations would decay gradually, so forecasters would be wise to consider both AR models and ARMA models for employment data.

We could begin with a base AR(2) model that adds in a constant value (μ) if all other values are zero. This is shown in the following generic formula:

$$y_t = \mu + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \varepsilon_t$$

Applying actual coefficients, our real AR(2) model might look something like:

$$y_t = 101.2413 + 1.4388 y_{t-1} - 0.4765 y_{t-2} + \varepsilon_t$$

We could also try to forecast our seasonally impacted employment data with an ARMA(3,1) model, which might look like the following formula:

$$y_t = \mu + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_3 y_{t-3} + \theta\varepsilon_{t-1} + \varepsilon_t$$

Applying actual coefficients, our real ARMA(3,1) model might look something like:

$$y_t = 101.1378 + 0.5004 y_{t-1} + 0.8722 y_{t-2} - 0.4434 y_{t-3} + 0.9709\varepsilon_{t-1} + \varepsilon_t$$

In practice, researchers would attempt to determine whether the AR(2) model or the ARMA(3,1) model provides a better prediction for the seasonally impacted data series.

Suppose the researcher settles on ARMA (3,1) model, and suppose the previous three values of the time series are as follows: $y_{t_1} = 10.38$, $y_{t_2} = 10.14$, $y_{t_3} = 10.50$, and suppose that previous shock $\varepsilon_{t_1} = -1.23$. The forecasted next period value of *y* would be calculated as:

$$y_t = 101.1378 + (0.5004 \times 10.38) + (0.8722 \times 10.14) - (0.4434 \times 10.50)$$

$$+ (0.9709 \times -1.23) = 109.3262$$

Sample and Partial Autocorrelations

LO 21.j: Describe sample autocorrelation and partial autocorrelation.

Sample autocorrelation and **partial autocorrelation** are calculated as discussed previously, but using sample data. These are used to validate and improve ARMA models. Initially, these sample statistics guide the analyst in selecting an appropriate model that conforms to the sample data.

Similarly, in evaluating the goodness of fit of a model, *residual* autocorrelations at different lags are computed. These are then tested for statistical significance. If the model fits the sample data well, none of the residual autocorrelations should be statistically significantly different from zero. In other words, we are trying to determine whether the model has captured all the information, or whether some information is still present in the residuals. We discuss formal tests for this in the next section.

Testing Autocorrelations

LO 21.k: Describe the Box-Pierce Q statistic and the Ljung-Box Q statistic.

As stated before, model specification checks involve an examination of residual ACF. We want all residual autocorrelations to be zero. A graphical examination of these autocorrelations can provide insights; any autocorrelations violating the 95% confidence interval around zero would indicate that the model does not adequately capture the underlying patterns in the data.

A joint test of determining that all residual autocorrelations equal zero versus at least one is not equal to zero is the **Box-Pierce (BP) statistic**:

$$Q_{BP} = T\sum_{i=1}^{h} r_i^2$$

where:
Q_{BP} = chi-squared statistic with *h* degrees of freedom
T = sample size

r_i = sample autocorrelation at lag i

For smaller samples (T ≤ 100), a version of the BP statistic known as the **Ljung-Box (LB) statistic** works better:

$$Q_{LB} = T\sum_{i=1}^{h}\left(\frac{T+2}{T-i}\right)r_i^2$$

Modeling Seasonality in an ARMA

LO 21.n: Explain how seasonality is modeled in a covariance-stationary ARMA.

Seasonality in time series data is evidenced by the recurrence of a pattern at the same time every year (e.g., higher retail sales in the fourth quarter). For a pure AR process, seasonality can be modeled by including a lag corresponding to the seasonality (i.e., fourth lag for quarterly data, twelfth lag for monthly data) in addition to any other relevant short-term lags. A similar approach is used for MA processes.

An ARMA model with seasonality is denoted by ARMA (p, q) × (p_S, q_S), where p_S and q_S denote the seasonal component. In this context, p_S and q_S are restricted to values of 1 or 0 (i.e., true or false) such that a value of 1 corresponds to the seasonal lag (e.g., 12 for monthly time series).

MODULE QUIZ 21.3

1. Which of the following statements about an autoregressive moving average (ARMA) process is correct?
   I. It involves autocorrelations that decay gradually.
   II. It combines the lagged unobservable random shock of the MA process with the observed lagged time series of the AR process.

   A. I only.
   B. II only.
   C. Both I and II.
   D. Neither I nor II.

2. Which of the following statements is correct regarding the usefulness of an autoregressive (AR) process and an autoregressive moving average (ARMA) process when modeling seasonal data?
   I. They both include lagged terms and, therefore, can better capture a relationship in motion.
   II. They both specialize in capturing only the random movements in time series data.

   A. I only.
   B. II only.
   C. Both I and II.
   D. Neither I nor II.

3. To test the hypothesis that the autocorrelations of a time series are jointly equal to zero based on a small sample, an analyst should most appropriately calculate:
   A. a Ljung-Box (LB) Q-statistic.
   B. a Box-Pierce (BP) Q-statistic.
   C. either a Ljung-Box (LB) or a Box-Pierce (BP) Q-statistic.
   D. neither a Ljung-Box (LB) nor a Box-Pierce (BP) Q-statistic.

KEY CONCEPTS

LO 21.a

To be covariance stationary, a time series must exhibit the following three properties:

1. Its mean must be stable over time.
2. Its variance must be finite and stable over time.
3. Its covariance structure must be stable over time.

LO 21.b

The covariance between the current value of a time series and its value τ periods in the past is referred to as its autocovariance at lag τ. Its autocovariances for all τ make up its autocovariance function. If a time series is covariance stationary, its autocovariance function is stable over time.

LO 21.c

White noise is a serially uncorrelated series with a mean of zero and a constant variance. If the observations in a white noise process are independent and uncorrelated, the process is referred to as independent white noise. If the process also follows a normal distribution, it is known as normal white noise or Gaussian white noise.

LO 21.d, 21.g, 21.m

An autoregressive (AR) process is specified in the form of a variable regressed against itself in lagged form. An AR(p) process is specified as:

$$y_t = d + \Phi y_{t-1} + \Phi y_{t-1} + \ldots + \Phi y_{t-p} + \varepsilon_t$$

Where the absolute values of all Φ coefficients should be less than one. The long-run (or unconditional) mean reverting level of an AR(p) series is computed as:

$$\mu = \frac{d}{1 - \Phi_1 - \Phi_2 \ldots - \Phi_p}$$

LO 21.e

A moving average (MA) process is a linear regression of the current values of a time series against both the current and previous unobserved white noise error terms, which are random shocks. MAs are always covariance stationary.

LO 21.f

A lag operator when applied to a value of a time series yields its lagged value:

$$y_{t-1} = Ly_t$$

An autoregressive (AR) process is covariance stationary only if its lag polynomial is invertible. This invertibility is used in the Box-Jenkins methodology to select the appropriate time series model.

LO 21.h

Autoregressive moving average (ARMA) models are used for those time series that show signs of both autoregressive (AR) and moving average (MA) processes. An ARMA(p,q) indicates *p* lags in the AR process and *q* lags in the MA process.

LO 21.i

If an autocorrelation plot for a data series cuts off abruptly, the forecaster should consider using an MA process. If the autocorrelations instead decay gradually, the forecaster should consider specifying either an autoregressive (AR) process or an autoregressive moving average (ARMA) process.

LO 21.j

Sample autocorrelations and partial autocorrelations are calculated using sample data and are used to validate and improve autoregressive moving average (ARMA) models. Initially, these sample statistics guide the analyst in selecting an appropriate model that conforms to the sample data. Residual autocorrelations at different lags are tested for statistical significance. If the model fits the sample data well, none of the residual autocorrelations should be statistically significantly different from zero.

LO 21.k

A joint test of determining that all residual autocorrelations equal zero versus at least one is not equal to zero is the Box-Pierce (BP) statistic:

$$Q_{BP} = T\sum_{i=1}^{h} r_i^2$$

where:
Q_{BP} = chi-squared statistic with *h* degrees of freedom
T = sample size
r_i = sample autocorrelation at lag *i*

For smaller samples (T ≤ 100), a version of the BP statistic known as the Ljung-Box (LB) statistic works better:

$$Q_{LB} = T\sum_{i=1}^{h}\left(\frac{T+2}{T-i}\right)r_i^2$$

LO 21.l

Both autoregressive (AR) and autoregressive moving average (ARMA) processes can be applied to time series data that show signs of seasonality. For example, we can forecast seasonally impacted data with an ARMA(3,1) model using the following formula:

$$y_t = \mu + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_3 y_{t-3} + \theta\varepsilon_{t-1} + \varepsilon_t$$

LO 21.n

Seasonality in time series data is evidenced by the recurrence of a pattern at the same time every year (e.g., higher retail sales in the fourth quarter). For a pure autoregressive (AR) process, seasonality can be modeled by including a lag corresponding to the seasonality (i.e., fourth lag for quarterly data, twelfth lag for monthly data) in addition to any other relevant short-term lags. A similar approach is used for moving average (MA) processes.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 21.1

1. **C** In theory, a time series can be infinite in length and still be covariance stationary. To be covariance stationary, a time series must have a stable mean, a stable

covariance structure (i.e., autocovariances depend only on displacement, not on time), and a finite variance. (LO 21.a)

2. **B** One feature that all ACFs have in common is that autocorrelations approach zero as the number of lags or displacements gets large. (LO 21.b)

3. **B** If a white noise process is Gaussian (i.e., normally distributed), it follows that the process is independent white noise. However, the reverse is not true; there can be independent white noise processes that are not normally distributed. Only those serially uncorrelated processes that have a zero mean and constant variance are white noise. (LO 21.c)

Module Quiz 21.2

1. **D** In order for an AR process to be covariance stationary, the sum of each of the slope coefficients should be less than 1. (LO 21.d)

2. **A** A key difference between an MA representation and an AR process is that the MA process shows autocorrelation cutoff while an AR process shows a gradual decay in autocorrelations. (LO 21.e)

3. **B** The coefficient is equal to 0.75, so using the concept derived from the Yule-Walker equation, the first-period autocorrelation is 0.75 (i.e., 0.75^1), and the second-period autocorrelation is 0.5625 (i.e., 0.75^2). (LO 21.e)

4. **B** There are two main purposes of using a lag operator. First, an AR process is covariance stationary only if its lag polynomial is invertible. Second, this invertibility is used in the Box-Jenkins methodology to select the appropriate time series model. (LO 21.f)

Module Quiz 21.3

1. **C** The ARMA process is important because its autocorrelations decay gradually and because it captures a more robust picture of a variable being estimated by including both lagged random shocks and lagged observations of the variable being estimated. The ARMA model merges the lagged random shocks from the MA process and the lagged time series variables from the AR process. (LO 21.h)

2. **A** Both AR models and ARMA models are good at forecasting with seasonal patterns because they both involve lagged observable variables, which are best for capturing a relationship in motion. It is the moving average representation that is best at capturing only random movements. (LO 21.i)

3. **A** The LB Q-statistic is appropriate for testing this hypothesis based on a small sample. (LO 21.k)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 11.

READING 22

NON-STATIONARY TIME SERIES

Study Session 7

EXAM FOCUS

The previous reading introduced methods to forecast a covariance stationary time series. Next, we will address non-stationary time series. Sources of non-stationarity fall into three main categories: time trends, seasonality, and unit roots (random walks). For the exam, be prepared to distinguish among these sources and identify the recommended approach to resolving them. Series with time trends can often be transformed into stationary series by estimating and removing the trend component. Seasonality can be modeled with dummy variables or by analyzing year-on-year changes. Time series with unit roots should be analyzed in terms of their change from the previous period.

MODULE 22.1: TIME TRENDS

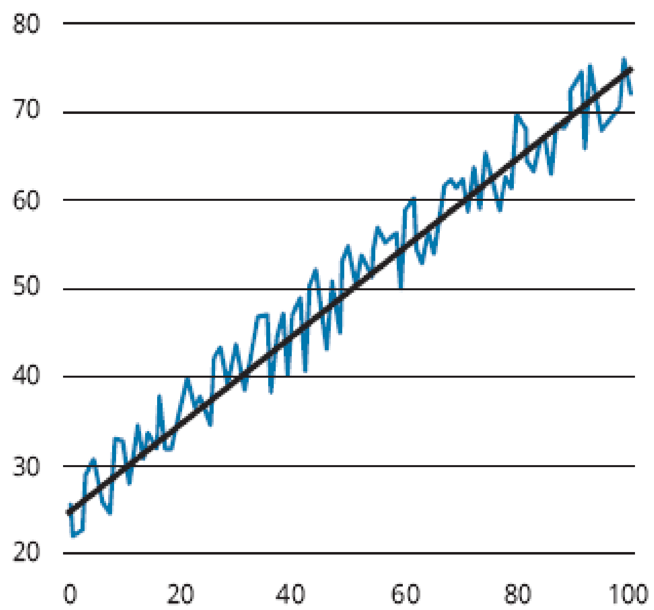
LO 22.a: Describe linear and nonlinear time trends.

LO 22.g: Calculate the estimated trend value and form an interval forecast for a time series.

Non-stationary time series may exhibit deterministic trends, stochastic trends, or both. **Deterministic trends** include both time trends and deterministic seasonality (which we will address in Module 22.2). **Stochastic trends** include *unit root* processes such as random walks (which we will address in Module 22.3).

Time trends may be linear or nonlinear. A series that exhibits a **linear time trend** is one that tends to change by the same amount each period. Graphically, such a series resembles deviations around an increasing or decreasing straight line, as in Figure 22.1.

Figure 22.1: Linear Time Trend



A linear time trend can be modeled simply as $y_t = \delta_0 + \delta_1 t + \varepsilon_t$, where ε_t is a white noise process. Note that what makes the series non-stationary is that the observations depend on time.

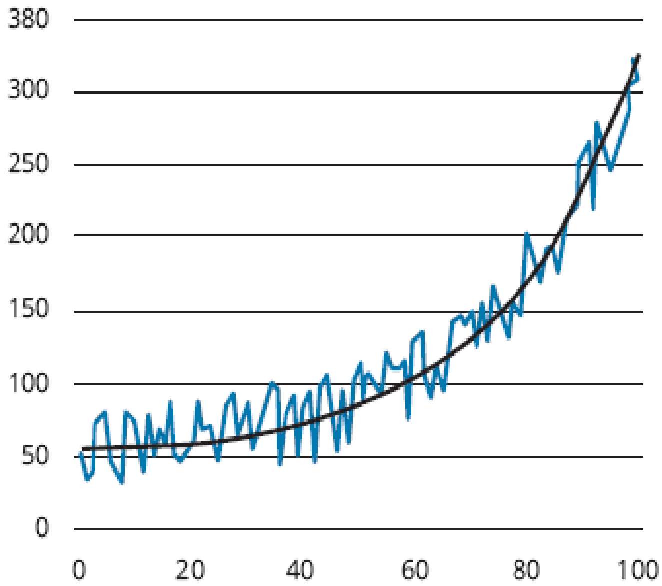
While linear time trend models benefit from simplicity, they are of limited use in finance and economics for two main reasons:

1. If the trend is downward, a linear model eventually produces negative values, which do not make sense when modeling quantities or prices.
2. Even if the trend is upward, a constant increase in the amount implies a decreasing rate of growth over time. Many variables are more accurately modeled as growing at a constant rate, rather than by a constant amount.

Fortunately, modeling techniques are not limited to linear time trends. An example of a **nonlinear time trend** (or **polynomial time trend**) model is a second-degree or quadratic polynomial model: $y_t = \delta_0 + \delta_1 + \delta_2 t^2 + \varepsilon_t$. Higher-order polynomials can also be modeled.

Many processes in finance and economics can be modeled using a **log-linear model**. A log-linear time trend represents a constant growth rate in a variable. This type of model is stated as $\ln(y_t) = \delta_0 + \delta_1 t + \varepsilon_t$. Graphically, they resemble the time series shown in Figure 22.2.

Figure 22.2: Log-Linear Time Trend



As with linear models, log-linear models can be extended to include polynomials, such as a **log-quadratic model**: $\ln(y_t) = \delta_0 + \delta_1 t + \delta_2 t^2 + \varepsilon_t$.

For a linear or nonlinear model, the trend value can be estimated by regression, as long as ε_t is white noise. If this assumption does not hold, a regression will produce misleading indicators of significance (the *t*-statistics of the coefficients) and goodness of fit (the R^2 of the regression), and a trend model alone is not sufficient to describe the time series.

Once we have estimated a model, we can use it to make forecasts. For example, with a simple linear trend model $y_t = \delta_0 + \delta_1 t + \varepsilon_t$, the forecast for one period ahead $(t + 1)$ would be: $y_{t+1} = \delta_0 + \delta_1(t + 1) + \varepsilon_{t+1}$, and a forecast for *h* periods ahead would be $y_{t+h} = \delta_0 + \delta_1(t + h) + \varepsilon_{t+h}$. With a logarithmic model, forecasting the level of a time series requires us to assume ε_t is normally distributed.

We can also use the regression results to place a confidence interval around a forecast. For example, assuming ε_t is normally distributed white noise, a 95% confidence interval for the forecast *h* periods ahead is $y_{t+h} \pm 1.96 \times$ standard deviation of the regression residuals.

Modeling and removing the trend component results in a *detrended* time series that we may be able to analyze further. Often the detrended time series is covariance stationary but not white noise. If so, we can improve on a trend model by using autoregressive (AR), moving average (MA), or autoregressive moving average (ARMA) techniques to forecast the detrended time series. (We described these techniques in Reading 21.)

MODULE QUIZ 22.1

1. An analyst has determined that monthly vehicle sales in the United States have been increasing over the last 10 years, but the growth rate over that period has been relatively constant. Which model is most appropriate to predict future vehicle sales?
   A. Linear model.
   B. Quadratic model.
   C. Log-linear model.
   D. Log-quadratic model.

2. Using data from 2001 to 2020, an analyst estimates a model for an industry's annual output as $Output_t = 80.163 + 4.248t + \varepsilon_t$, from a regression with a residual standard deviation of 107.574. Assume *t* equals a given full year (e.g., 2021) and that the error term is normally distributed. A 95% confidence interval for a forecast of 2021 industry output is closest to:
   A. 8,374 to 8,796.
   B. 8,455 to 8,876.
   C. 8,477 to 8,693.
   D. 8,557 to 8,773.

MODULE 22.2: SEASONALITY

LO 22.b: Explain how to use regression analysis to model seasonality.

Seasonality in a time series is a pattern that tends to repeat from year to year. One example is monthly sales data for a retailer. Because sales data normally varies according to the calendar, we might expect this month's sales (x_t) to be related to sales for the same month last year (x_{t-12}).

Specific examples of seasonality relate to increases that occur at only certain times of the year. For example, purchases of retail goods typically increase dramatically every year in the weeks leading up to Christmas. Similarly, sales of gasoline generally increase during the summer months when people take more vacations. Weather is another common example of a seasonal factor as production of agricultural commodities is heavily influenced by changing seasons and temperatures.

Seasonality in a time series can also refer to cycles shorter than a year. For example, a daily time series may exhibit deterministic effects on a specific day of the week. We use the more general term **calendar effects** to refer to any cycles that may recur within a year or less.

An effective technique for modeling seasonality is to include seasonal **dummy variables** in a regression. Seasonal dummy variables can take a value of either *one* or *zero* to represent the season being *on* or *off*. For example, in a time series regression of monthly stock returns, we might incorporate a *January* dummy variable that would take on the value of *one* if a stock return occurred in January, and *zero* if it occurred in any other month. The January dummy variable helps us to see if stock returns in January were significantly different than stock returns in all other months of the year. Many *January effect* anomaly studies use this type of regression methodology.

A regression model can include dummy variables for up to one less than the frequency of the data. For example, a model for quarterly time series can have up to three seasonal dummy variables, and a model for a monthly time series can have up to 11. The "extra" period is accounted for by the condition that all the other dummy variables equal zero. (If we included a dummy variable for the fourth quarter or the 12th month, we would bring multicollinearity into our regression because the value of one dummy variable could be predicted exactly from the values of the others.)

Another approach to modeling seasonality is **seasonal differencing**. Instead of modeling the level of a series, we can model the differences between its level and its

year-ago level. Seasonal differencing can also help in modeling series with time trends and unit roots.

EXAMPLE: Seasonal dummy variables

Consider the following regression equation for explaining quarterly earnings per share (EPS) in terms of the quarter of their occurrence:

$$EPS_t = \beta_0 + \beta_1 D_{1,t} + \beta_2 D_{2,t} + \beta_3 D_{3,t} + \varepsilon_t$$

where:
EPS_t = a quarterly observation of earnings per share
$D_{1,t}$ = 1 if period *t* is the first quarter of a year, $D_{1,t}$ = 0 otherwise
$D_{2,t}$ = 1 if period *t* is the second quarter of a year, $D_{2,t}$ = 0 otherwise
$D_{3,t}$ = 1 if period *t* is the third quarter of a year, $D_{3,t}$ = 0 otherwise

The intercept term, β_0, represents the average value of EPS for the fourth quarter. The slope coefficient on each dummy variable estimates the *difference* in EPS (on average) between the respective quarter (i.e., quarter one, two, or three) and the omitted quarter (the fourth quarter, in this case). Think of the omitted class as the reference point.

Suppose we estimate the quarterly EPS regression model with 10 years of data (40 quarterly observations) and find that β_0 = 1.25, β_1 = 0.75, β_2 = −0.20, and β_3 = 0.10:

$$\widehat{EPS}_t = 1.25 + 0.75D_{1,t} - 0.20D_{2,t} + 0.10D_{3,t}$$

Determine the average EPS in each quarter over the past 10 years.

Answer:

The average EPS in each quarter over the past 10 years is as follows:

- average fourth-quarter EPS = 1.25
- average first-quarter EPS = 1.25 + 0.75 = 2.00
- average second-quarter EPS = 1.25 − 0.20 = 1.05
- average third-quarter EPS = 1.25 + 0.10 = 1.35

These are also the model's predictions of future EPS in each quarter of the following year. For example, to use the model to predict EPS in the first quarter of the next year, set $\hat{D}_{1,t} = 1$, $\hat{D}_{2,t} = 0$, and $\hat{D}_{3,t} = 0$. Then $\widehat{EPS}_t$ = 1.25 + 0.75(1) − 0.20(0) + 0.10(0) = 2.00. This simple model uses average EPS for a specific quarter over the past 10 years as the forecast of EPS in its respective quarter of the following year.

LO 22.f: Explain how to construct an h-step-ahead point forecast for a time series with seasonality.

Forecasting a seasonal series is fairly straightforward. A pure seasonal dummy variable model can be constructed as follows:

$$y_t = \sum_{i=1}^{s} \gamma_i (D_{i,t}) + \varepsilon_t$$

After adding a time trend, the model can then take the following form:

$$y_t = \beta_1(t) + \sum_{i=1}^{s}\gamma_i\left(D_{i,t}\right) + \varepsilon_t$$

We can expand the forecasting model even further by allowing for other calendar effects. For example, if we suspect a time series exhibits holiday variations (HDV) and trading-day variations, we can account for them with additional dummy variables:

$$y_t = \beta_1(t) + \sum_{i=1}^{s}\gamma_i\left(D_{i,t}\right) + \varepsilon_t + \sum_{i=1}^{v_1}\delta_i^{HDV}\left(HDV_{i,t}\right) + \sum_{i=1}^{v_2}\delta_i^{TDV}\left(TDV_{i,t}\right) + \varepsilon_t$$

This complete model can now be used for *out-of-sample* forecasts at time T + h by constructing an **h-step-ahead point forecast** as follows:

$$y_{T+h} = \beta_1(T+h) + \sum_{i=1}^{s}\gamma_i\left(D_{i,T+h}\right) + \sum_{i=1}^{v_1}\delta_i^{HDV}\left(HDV_{i,T+h}\right) + \sum_{i=1}^{v_2}\delta_i^{TDV}\left(TDV_{i,T+h}\right) + \varepsilon_{T+h}$$

That is, determine the value for the time trend at time T + h and set the dummy variables to their appropriate 0 or 1 values for the period T + h.

MODULE QUIZ 22.2

1. Jill Williams is an analyst in the retail industry. She is modeling a company's sales and has noticed a quarterly seasonal pattern. If Williams includes an intercept term in her model, how many dummy variables should she use to model the seasonality component?
   A. 1.
   B. 2.
   C. 3.
   D. 4.
2. Consider the following regression equation utilizing dummy variables for explaining quarterly EPS in terms of the quarter of their occurrence:

   $EPS_t = \beta_0 + \beta_1 D_{1,t} + \beta_2 D_{2,t} + \beta_3 D_{3,t} + \varepsilon_t$

   where:

   EPS_t = a quarterly observation of EPS

   $D_{1,t}$ = 1 if period *t* is the first quarter, $D_{1,t}$ = 0 otherwise

   $D_{2,t}$ = 1 if period *t* is the second quarter, $D_{2,t}$ = 0 otherwise

   $D_{3,t}$ = 1 if period *t* is the third quarter, $D_{3,t}$ = 0 otherwise

   The intercept term β_0 represents the average value of EPS for the:
   A. first quarter.
   B. second quarter.
   C. third quarter.
   D. fourth quarter.
3. A model for the change in a retailer's quarterly sales, using seasonal dummy variables DQ, is estimated as:

   $\Delta Sales_t = 4.9 - 2.1D_{Q2} - 3.8D_{Q3} + 6.5D_{Q4}$

   In the third quarter, sales are forecast to:
   A. decrease by 3.8.
   B. decrease by 1.0.
   C. increase by 1.1.
   D. increase by 3.8.

MODULE 22.3: UNIT ROOTS

LO 22.c: Describe a random walk and a unit root.

We describe a time series as a **random walk** if its value in any given period is its previous value plus-or-minus a random “shock.” Symbolically, we state this as $y_t = y_{t-1} + \varepsilon_t$.

This seems simple enough, but if $y_t = y_{t-1} + \varepsilon_t$, it follows logically that the same was true in earlier periods, $y_{t-1} = y_{t-2} + \varepsilon_{t-1}$, $y_{t-2} = y_{t-3} + \varepsilon_{t-2}$ and so forth, all the way back to the beginning of the time series: $y_1 = y_0 + \varepsilon_1$.

If we substitute these (recursively) back into $y_t = y_{t-1} + \varepsilon_t$, we eventually get: $y_t = y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t$. That is, any observation in the series is a function of the beginning value and all the past shocks, as well as the shock in the observation’s own period.

A key property of a random walk is that its variance increases with time. This implies a random walk is not covariance stationary, so we cannot model one directly with AR, MA, or ARMA techniques.

A random walk is a special case of a wider class of time series known as **unit root** processes. They are called this because when expressed using lag polynomials (which we introduced in Reading 21), one of their roots is equal to 1, as in: $(1 - L)(1 - 0.65L)\, y_t = \varepsilon_t$.

PROFESSOR’S NOTE

A unit root process is sometimes described as a *random walk with drift*. For our purposes here, we can think of *random walk* and *unit root* more-or-less interchangeably.

LO 22.d: Explain the challenges of modeling time series containing unit roots.

If we attempt to model a time series directly when it has a unit root, we run into three main problems:

1. Unlike stationary time series, a series with a unit root does not revert to a mean.
2. Time series with unit roots often show spurious relationships with each other.
3. If we use an ARMA model, its estimated parameters follow an asymmetric distribution that depends on the sample size and the presence of a time trend (a **Dickey-Fuller distribution**). This reduces our ability to select a correct model or make valid forecasts.

All of these problems can be addressed by modeling the **first differences** of a unit root time series, which is to say their changes from one period to the next. In fact, modeling first differences also can address time trends and seasonality.

If the first differences are not a stationary series, we can take the differences of those (i.e., double differencing). However, if we already have a stationary series, taking its differences results in **overdifferencing**, which adds complexity to a forecasting model instead of reducing its complexity.

LO 22.e: Describe how to test if a time series contains a unit root.

The most common way to test a series for a unit root is with an **augmented Dickey-Fuller test**. This is essentially a test of whether the lagged level of a series is a statistically significant factor in a regression model. That model can also include deterministic factors and lagged differences of the series, as appropriate. Such a model may be stated as:

$$\Delta Y_t = \gamma Y_{t-1} + \delta_0 + \delta_1 t + \lambda_1 \Delta Y_{t-1} + \ldots + \lambda_p \Delta Y_{t-p} + \varepsilon_t$$

where the δs represent deterministic factors and the λs represent lagged differences. The model should include just enough of these to make ε_t a white noise process.

The null hypothesis is that γ, the coefficient on the lagged value Y_{t-1}, is equal to zero. If we fail to reject the hypothesis, the lagged level of the series has no predictive value and the series is a random walk. If the series is covariance stationary, then γ will be significantly less than zero. (If γ is significantly greater than zero, the series is not covariance stationary because it's an explosive process rather than a random walk.) So although the null hypothesis is $\gamma = 0$, the alternative hypothesis is $\gamma < 0$, and not $\gamma \neq 0$.

MODULE QUIZ 22.3

1. A random walk is most accurately described as a time series whose value is a function of its:
   A. previous value only.
   B. beginning value only.
   C. previous value and a random shock.
   D. beginning value and all historical shocks.

2. An augmented Dickey-Fuller test will reject the hypothesis that a process is a unit root if the coefficient on the lagged value is statistically significantly:
   A. less than zero.
   B. equal to zero.
   C. greater than zero.
   D. different from zero.

KEY CONCEPTS

LO 22.a

A time series that tends to grow by a constant amount each period has a linear trend. A time series that tends to grow at a constant rate each period has a nonlinear trend.

LO 22.b

A regression model can account for seasonality by introducing dummy variables to represent seasonal effects. To avoid multicollinearity, the number of dummy variables

must be one less than the number of periods in a year (e.g., three dummy variables for quarterly data).

LO 22.c

A time series is a random walk if its value in any given period is its previous value plus-or-minus a random "shock." A random walk is not covariance stationary. Random walks are a special case of a wider class known as unit root processes, called this because when expressed using lag polynomials, one of their roots is equal to one.

LO 22.d

Unlike stationary time series, a series with a unit root does not revert to a mean. Time series with unit roots often show spurious relationships. Model parameters for a unit root series follow a Dickey-Fuller distribution, which reduces our ability to select a correct model or make valid forecasts. All of these problems can be addressed by modeling the differences of the unit root series.

LO 22.e

The most common way to test a series for a unit root is with an augmented Dickey-Fuller test. This is a test of whether the lagged level is a statistically significant factor in a regression model. The null hypothesis is that the coefficient on the lagged value is equal to zero. The alternative hypothesis, however, is that the coefficient is less than zero, not different from zero.

LO 22.f

Given a model with a time trend and seasonal dummy variables, we can construct an h-step-ahead point forecast by determining the value of the time trend at time T + h and setting the dummy variables to their appropriate 0 or 1 values for the period T + h.

LO 22.g

Assuming a model's forecast error is a normally distributed white noise process, a 95% confidence interval for the forecast *h* periods ahead is y_{t+h} ± 1.96 × standard deviation of the regression residuals.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 22.1

1. **C** A log-linear model is most appropriate for a time series that grows at a relatively constant growth rate. (LO 22.a)

2. **B** For t = 2021, a point forecast for industry output is 80.163 + 4.248(2021) = 8,665.371. A 95% confidence interval is 8,665.371 ± 1.96(107.574) = 8,454.526 to 8,876.216. (LO 22.g)

Module Quiz 22.2

1. **C** Whenever we want to distinguish between *s* seasons in a model that incorporates an intercept, we must use s − 1 dummy variables. For example, if we have quarterly data, s = 4, and thus we would include s − 1 = 3 seasonal dummy variables. (LO 22.b)

2. **D** The intercept term represents the average value of EPS for the fourth quarter. The slope coefficient on each dummy variable estimates the difference in EPS (on average) between the respective quarter (i.e., quarter one, two, or three) and the omitted quarter (the fourth quarter, in this case). (LO 22.b)

3. **C** $\Delta Sales_{Q3} = 4.9 - 2.1(0) - 3.8(1) + 6.5(0) = 1.1$ (LO 22.f)

Module Quiz 22.3

1. **D** For a random walk, $y_t = y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t$, so its value at time *t* is a function of its beginning value and all shocks, as well as the shock in the observation's own period. (LO 22.c)

2. **A** Although the null hypothesis is that the coefficient on the lagged value is equal to zero, the rejection condition is that the coefficient is less than zero. (LO 22.e)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 12.

READING 23

MEASURING RETURN, VOLATILITY, AND CORRELATION

Study Session 7

EXAM FOCUS

Traditionally, volatility has been synonymous with risk. Thus, the accurate estimation of volatility is crucial to understanding potential risk exposure. For the exam, understand how to calculate simple and continuously compounded returns and recognize differences between definitions of volatility. Since financial returns tend to follow nonnormal distributions, it is important to understand the properties of this distribution, how to test for this type of distribution, and what the tails look like in this distribution. This reading closes with the concepts of correlations and dependence and how to test for them using various methods.

MODULE 23.1: DEFINING RETURNS AND VOLATILITY

Simple and Continuously Compounded Returns

LO 23.a: Calculate, distinguish, and convert between simple and continuously compounded returns.

Returns on investments are often expressed as simple returns and continuously compounded returns. A **simple return** can be expressed over various periods of time, spanning from a single hour to a full year. Across multiple time periods, an asset's return can be calculated by taking the product of each period's simple return.

Assuming an asset (priced at P) is purchased at time t − 1 and sold at time t, the simple return (R) is equal to:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Continuously compounded (log) returns (r) can be calculated using the following formula:

$r_t = \ln P_t - \ln P_{t-1}$

For log returns, summing single period log returns produces a multiple period total return with the following formula:

$$r_T = \sum_{t=1}^{T} r_t$$

For shorter time horizons, log returns are more appropriate than simple returns. Log returns also do not accurately approximate simple returns when the simple return is large.

The following equation can be used to convert between the simple (R) and log return (r), with the simple return always exceeding the log return:

$1 + R_t = \exp r_t$

Volatility, Variance, and Implied Volatility

LO 23.b: Define and distinguish among volatility, variance rate, and implied volatility.

The **volatility** of a variable, σ, is expressed as the standard deviation of its returns. The variance (or **variance rate**) of an asset is expressed as σ^2. Volatility, along with the mean (μ) and a shock (e_t) with a mean of zero and a variance of one, can be used to derive the return (r) on an asset using the following formula:

$r_t = \mu + \sigma e_t$

For basic modeling, the return calculated across multiple periods is the sum of the individual returns. The mean of a weekly return, assuming returns are calculated daily, is 5μ; the weekly variance is $5\sigma^2$, and the weekly volatility is $\sqrt{5}\sigma$. The annualized volatility (using monthly returns to measure volatility) is calculated as:

$$\sigma_{annual} = \sqrt{12 \times \sigma^2_{monthly}}$$

Assuming 252 trading days in the calendar year and volatility measured daily, the annualized volatility is calculated as:

$$\sigma_{annual} = \sqrt{252 \times \sigma^2_{daily}}$$

Options are used to calculate **implied volatility**, which is an annual volatility number that can be measured by backing into it using option prices. The Black-Scholes-Merton (BSM) model used to calculate the price of a call option includes inputs for the current asset price, strike price, time to maturity, risk-free interest rate, and annual variance. As long as the option price is known, the other variables except the annual variance are all observable and can be used to back into the variance number. The model's inherent assumption that variance is constant over time is one drawback to using this approach to calculate implied volatility.

The VIX Index is used to measure implied volatility for the S&P 500 for a prospective period covering 30 calendar days. The methodology uses option prices with future expiration dates and multiple strike prices, and, therefore, serves as a forward-looking volatility measure. The VIX method, which requires a significant and liquid derivatives market, can be computed for assets like equity indices, U.S. Treasury bonds, commodities, and individual stocks.

MODULE QUIZ 23.1

1. Assuming a simple return of 5.00%, the log return will be closest to:
   A. 4.88%.
   B. 5.00%.
   C. 5.05%.
   D. 5.13%.

2. Which of the following statements is correct in regard to using the Black-Scholes-Merton (BSM) pricing model to calculate implied volatility?
   A. The option price is not needed for the calculation.
   B. Variance is assumed to remain constant over time.
   C. Time to maturity is not one of the components of the calculation.
   D. The current asset price has to remain constant in the calculation.

MODULE 23.2: NORMAL AND NONNORMAL DISTRIBUTIONS

LO 23.c: Describe how the first two moments may be insufficient to describe non-normal distributions.

The first two moments of a probability density function are its mean and variance, respectively. The third moment is skewness and the fourth moment is kurtosis. For a normal distribution, which has thin tails and is symmetric, there is no skewness or excess kurtosis. However, the reality is that financial returns often follow a nonnormal distribution, so there is skewness and excess kurtosis.

When examining returns for the S&P 500, the Japanese Yen (JPY)/U.S. Dollar (USD) exchange rate, and gold over a period of time, each of these assets exhibits skewness that is not equal to zero and each exhibits kurtosis larger than three (implying positive excess kurtosis). For the first two assets, the skewness is negative while gold returns reflect positive skewness.

Jarque-Bera Test

LO 23.d: Explain how the Jarque-Bera test is used to determine whether returns are normally distributed.

The **Jarque-Bera (JB) test statistic** can be used to test whether a distribution is normal, meaning that there is zero skewness and no excess kurtosis ($K - 3 = 0$). If skewness is S and kurtosis is K, the null and alternative hypotheses are as follows:

Null:

H_0: S = 0 and K = 3

Alternative:

H_A: S ≠ 0 and K ≠ 3

The test statistic, where *T* is the sample size, is:

$$JB = (T-1)\left(\frac{\hat{S}^2}{6} + \frac{(\hat{K}-3)^2}{24}\right)$$

Because both the skewness and kurtosis components of the equation are asymptotically normally distributed and uncorrelated, each has a chi-squared distribution (χ_1^2) such that the JB is approximately χ_2^2. Smaller values will indicate that the null hypothesis is likely true, while larger values are indicative of a null that is likely to be rejected. At 5% and 1% respectively, and with critical values of 5.99 and 9.21, the null will be rejected if the JB calculation is above these levels. It is often the case that the longer the time period for measurement, the more likely it is that the JB statistic is smaller and the financial return approximates a normal distribution.

EXAMPLE: Test the hypothesis that a fund's returns follow a normal distribution

Assume 60 monthly returns are sampled for a fund with skewness equal to 0.30 (slight positive skewness) and kurtosis equal to 3.50. Using a 95% confidence interval, the critical value for the chi-squared distribution with two degrees of freedom equals 5.99 (i.e., corresponds to the 95th percentile of the chi-squared distribution).

Construct a test of the hypothesis that the fund's returns follow a normal distribution.

Answer:

The JB statistic for this fund equals:

$$JB = (59)\left(\frac{0.3^2}{6} + \frac{0.5^2}{24}\right)$$
$$= 59(0.015 + 0.0104) = 1.5$$

The JB statistic for this fund (1.5) is less than the critical value (5.99). The statistic does not lie in the rejection area to the right of 5.99. Therefore, we would fail to reject the hypothesis that this fund's returns follow a normal distribution.

The Power Law

LO 23.e: Describe the power law and its use for non-normal distributions.

As financial returns tend to follow nonnormal distributions, studying the tails can help explain how returns are distributed in reality. In a normal distribution with a kurtosis of three (and excess kurtosis of zero), the tails are thin. For other distributions, the tails do not decline as quickly. Some of these distributions (including the Student's *t*-distribution) have power law tails, implying that the probability of seeing a return larger than a specific value of *x* (with constants: k and α) is equal to: $P(X > x) = kx^{-\alpha}$. Fat tails with slow declines are found in power law tails and distributions like the Student's *t*-distribution, which explains why observations away from the mean are more common than those found in normal distributions.

MODULE QUIZ 23.2

1. Relative to a normal distribution, financial returns tend to have a nonnormal distribution, which will have:
   A. thin tails.
   B. kurtosis greater than three.
   C. minimal to no skewness.
   D. a symmetrical distribution.
2. Which of the following statements regarding the Jarque-Bera (JB) test statistic is most accurate?
   A. The null hypothesis states that skewness does not equal zero.
   B. The alternative hypothesis states that kurtosis is equal to three.
   C. The alternative hypothesis is likely to be rejected when the JB statistic is high.
   D. The null hypothesis is likely to not be rejected when the JB statistic is very small.
3. Which of the following statements is most accurate regarding power law tails?
   A. More observations tend to be closer to the mean.
   B. The standard normal distribution exhibits power law tails.
   C. The tails exhibit faster declines than normally distributed tails.
   D. They tend to have “fatter” tails than those found in a normal distribution.

MODULE 23.3: CORRELATIONS AND DEPENDENCE

LO 23.f: Define correlation and covariance and differentiate between correlation and dependence.

LO 23.g: Describe properties of correlations between normally distributed variables when using a one-factor model.

LO 23.h: Compare and contrast the different measures of correlation used to assess dependence.

Random variables can either be independent or dependent. If they are independent, the product of their marginal densities will equal their joint density per the equation: $f_{x,y}(x,y) = f_x(x)f_y(y)$.

Diversification benefits increase and tail risk decreases when variables are independent. However, financial assets tend to be highly dependent from both a linear and nonlinear perspective. **Pearson's correlation** serves as a method of measuring linear dependence. Correlation represents the linear relationship between two variables, while covariance represents the directional relationship between two variables.

Regression is the link tying correlation to linear dependence. In the standard regression equation $Y_i = \alpha + \beta X_i + \varepsilon_i$, if Y and X are standardized such that they each have a variance of one (termed, unit variance), the correlation will be equal to the regression slope (β).

When dependence is nonlinear, there is no one statistic used to measure it. To measure nonlinear dependence, measures such as **Spearman's rank correlation** and **Kendall's τ (tau)** can be used. The values for both must lie between −1 and 1, they are each zero when the returns are completely independent, they are scale invariant, and both are positive (negative) based on the increasing (decreasing) relationship between the variables.

Spearman's Rank Correlation

Spearman's correlation is a linear correlation estimator which is applied to ranks of observations. The strength of the linear relationship between ranks, as opposed to the linear relationship between the variables, drives rank correlation. In a situation where two random variables (X and Y) have *n* associated observations, $Rank_X$ and $Rank_Y$ serve as the ranks of the variables. *One* equates to the smallest value of each variable, *two* the second smallest, *three* the third smallest, and the trend continues on until *n* serves as the largest rank. The equation for the correlation estimator is:

$$\hat{\rho}_s = \frac{\widehat{Cov}[Rank_X, Rank_Y]}{\sqrt{\hat{V} Rank_X}\sqrt{\hat{V} Rank_Y}}$$

Assuming ranks are distinct, and $Rank_{Xi} - Rank_{Yi}$ represents the difference (d_i) in ranks for the same observation, the following equation can be used to express the estimator:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2 - 1)}$$

The correlation will be close to 1 when highly ranked values of X and Y are paired together. However, when the largest values of one variable are grouped with the smallest values of another, the variables will have strong negative dependence, the difference will be large, and correlation will be close to −1. If the variables themselves have a strong linear relationship, rank and linear correlation will be similar. If there are large differences in linear and rank correlations, there is likely to be a key nonlinear relationship. Rank correlation, unlike linear correlation, is not as sensitive or vulnerable to outliers because *ranks* rather than *variable values* are used.

EXAMPLE: **Spearman's rank correlation**

Calculate the Spearman rank correlation for the returns of stocks X and Y provided in the following table.

Returns for Stocks X and Y

Year	X	Y
2011	25.0%	–20.0%
2012	–20.0%	10.0%
2013	40.0%	20.0%
2014	–10.0%	30.0%

Answer:

The calculations for determining the Spearman rank correlation coefficient are shown in the table below. The first step involves ranking the returns for stock X from lowest to highest in the second column. The first column denotes the respective year for each return. The returns for stock Y are then listed for each respective year. The fourth and fifth columns rank the returns for variables X and Y. The differences between the rankings for each year are listed in column six. Lastly, the sum of squared differences in rankings is determined in column 7.

Ranking Returns for Stocks X and Y

Year	X	Y	X Rank	Y Rank	d_i	d_i^2
2012	–20.0%	10.0%	1	2	–1	1
2014	–10.0%	30.0%	2	4	–2	4
2011	25.0%	–20.0%	3	1	2	4
2013	40.0%	20.0%	4	3	1	1
					Sum =	10

The Spearman rank correlation coefficient can then be determined as 0.0:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 10}{4(16-1)} = 0.0$$

Kendall's τ

Kendall's $_\tau$ is used to measure *concordant* and *discordant* pairs and their relative frequency. The measure represents the difference between the probabilities of concordance and discordance.

To see how this is applied, take two random variables (X_i,Y_i) and (X_j,Y_j). If $(X_i < X_j)$ and $(Y_i < Y_j)$, the relative positions of X and Y are in agreement and the pair is concordant. If the orders are different, the pair will be discordant. If $X_i = X_j$ and $Y_i = Y_j$, the pair is neither concordant nor discordant. Random variables with many concordant pairs tend to have strong positive dependence, whereas variables with many discordant pairs tend to have a strong negative relationship.

The equation for calculating Kendall's τ is:

$$\hat{\tau} = \frac{n_c - n_d}{n(n-1)/2} = \frac{n_c}{n_c + n_d + n_t} - \frac{n_d}{n_c + n_d + n_t}$$

where:
n_c = number of concordant pairs
n_d = number of discordant pairs
n_t = number of ties

If all pairs are concordant, the output will equal exactly 1. If all pairs are discordant, the output will equal −1. Any other pattern will produce a number between −1 and 1.

EXAMPLE: **Kendall's τ**

Calculate the Kendall τ correlation coefficient for the stock returns of *X* and *Y* listed below.

Ranked Returns for Stocks X and Y

Year	X	Y	X Rank	Y Rank
2012	−20.0%	10.0%	1	2
2014	−10.0%	30.0%	2	4
2011	25.0%	−20.0%	3	1
2013	40.0%	20.0%	4	3

Answer:

Begin by comparing the rankings of *X* and *Y* stock returns in columns four and five of the table above. There are four pairs of observations, so there will be six combinations. The following table summarizes the pairs of rankings based on the stock returns for *X* and *Y*. There are three concordant pairs and three discordant pairs.

Categorizing Pairs of Stock X and Y Returns

Concordant Pairs	Discordant Pairs
{(1,2),(2,4)}	{(1,2),(3,1)}
{(1,2),(4,3)}	{(2,4),(3,1)}
{(3,1),(4,3)}	{(2,4),(4,3)}

Kendall's τ can then be determined as 0:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} = \frac{3-3}{4(4-1)/2} = 0$$

Thus, there is no positive or negative relationship between the stock returns of *X* and *Y* based on the Kendall τ correlation coefficient.

Positive Definiteness

When all random variables have unit variance (variance equal to 1), the correlation matrix and covariance matrix are the same thing. Every linear combination of random variables must have a variance which is non-negative. **Positive definiteness**, defined as every weighted average combination having a positive variance, requires that the variance of an average of components in a covariance matrix must be positive.

In order to ensure that correlation matrices are positive definite, two structured correlations are typically used. The first type, known as equicorrelation, sets all correlations equal to the same amount. The second type applies a structure which assumes that correlations are due to a common factor exposure, thereby making the correlation for any entries into the matrix equal to $\rho_{i,j} = Y_i Y_j$ and each entry having a correlation between −1 and 1.

MODULE QUIZ 23.3

1. An analyst calculates a Spearman's rank correlation of 0.48. This output is indicative of:
   A. positive linear correlation.
   B. negative linear correlation.
   C. positive nonlinear dependence.
   D. negative nonlinear dependence.
2. Which of the following situations is indicative of equicorrelation in a correlation matrix?
   A. Correlations which are all equal to 1.
   B. Variables with correlations other than 0.
   C. Variables with negative coefficients of determination.
   D. Three variables with a correlation with one another of 1.25.

KEY CONCEPTS

LO 23.a

Investment returns can be stated as both simple and continuously compounded (log) returns. With simple returns, an asset's return across multiple time periods is calculated by taking the product of each period's simple return. For continuously compounded returns, an asset's return across multiple time periods is calculated by taking the sum of each single period's log returns. Log returns, which are always less than simple returns, are more often used for shorter time horizons. The equation $1 + R_t = \exp r_t$ can be used to convert between simple (R) and log (r) returns.

LO 23.b

The volatility of a variable, σ, is expressed as the standard deviation of its returns. The variance (or variance rate) of an asset is expressed as σ^2. Options are used to calculate implied volatility, which is an annual volatility number that can be measured by backing into it using option prices. All of the variables included in the Black-Scholes-Merton (BSM) model used to calculate call option prices are observable except the annual variance, meaning that the variance value can be derived as long as the price of the option is known. The VIX Index is a forward-looking methodology used to measure implied volatility for the S&P 500 for a prospective period covering 30 calendar days.

LO 23.c

The first two moments of a probability density function are its mean and variance, which are used to describe a normal distribution. The third moment is the skewness and the fourth moment is the kurtosis. For a normal distribution, which has thin tails and is symmetric, there is no skewness or excess kurtosis. Financial returns often follow a nonnormal distribution and as such, there is skewness and excess kurtosis.

LO 23.d

The Jarque-Bera (JB) test statistic can be used to test whether a distribution is normal, meaning that there is zero skewness and no excess kurtosis (K – 3 = 0). If the result falls below the critical value, the null will not be rejected and the distribution will be deemed normal. If the result is above the critical value, the null will be rejected. Financial returns are likely to follow a more normal distribution over longer time periods.

LO 23.e

Normal distributions have a kurtosis of three (and excess kurtosis of zero) and thin tails. For other types of distributions, the tails do not decline as quickly. Fat tails with slow declines are found in power law tails and distributions like the Student's *t*-distribution, which explains why observations away from the mean are more common than those found in normal distributions.

LO 23.f and 23.h

Correlation represents the linear relationship between two variables, while covariance represents the directional relationship between two variables. Random variables can either be independent or dependent. Pearson's correlation serves as a method of measuring linear dependence.

When dependence is nonlinear, measures such as Spearman's rank correlation and Kendall's τ (tau) can be used. As with traditional correlation measures, the values for both must lie between –1 and 1.

Spearman's correlation is a linear correlation estimator which is applied to ranks of observations. The strength of the linear relationship between ranks, as opposed to the linear relationship between the variables, drives rank correlation. The correlation will be close to 1 when highly ranked values of X and Y are paired together. However, when the largest values of one variable are grouped with the smallest values of another, the variables will have strong negative dependence, the difference will be large, and correlation will be close to –1.

Kendall's τ is used to measure concordant and discordant pairs and their relative frequency. The measure represents the difference between the probabilities of concordance and discordance. If all pairs are concordant, the output will equal exactly 1. If all pairs are discordant, the output will equal –1. Any other pattern will produce a number between –1 and 1.

LO 23.g

Every linear combination of random variables must have a variance which is non-negative. Positive definiteness, defined as every weighted average combination having a positive variance, requires that the variance of an average of components in a covariance matrix must be positive. In order to ensure that correlation matrices are positive definite, two structured correlations are typically used. The first type, known as equicorrelation, sets all correlations equal to the same amount. The second type applies a structure which assumes that correlations are due to a common factor exposure.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 23.1

1. **A** The equation to convert the simple return to the log return is:

   $1 + R_t = \exp r_t$

   Plugging in values, 1.05 = exp r_t. Taking the natural log of each side to isolate the log return (r) results in ln 1.05 = 0.0488 or 4.88%. (LO 23.a)

2. **B** One of the drawbacks to using the BSM pricing model to derive implied volatility is that variance must remain constant over time. The option price and time to maturity are both needed for the calculation, but there is no requirement that the current underlying asset price has to remain constant. (LO 23.b)

Module Quiz 23.2

1. **B** A nonnormal distribution is likely to have either positive or negative skewness and a kurtosis that is different from three. A normal distribution has thin tails, kurtosis equal to three, no skewness, and a symmetrical distribution. (LO 23.c)

2. **D** When the JB test statistic is very small, the null hypothesis is likely to not be rejected. When the statistic is high, the null is likely to be rejected (with the alternative hypothesis not being rejected). The null hypothesis states that skewness is zero and kurtosis is three (with excess kurtosis therefore equal to zero). The alternative hypothesis states that skewness is not equal to zero and kurtosis is not equal to three. (LO 23.d)

3. **D** Power law tails tend to be "fatter" than the tails found in normal distributions. Power law tails reflect more observations found farther away from the mean and they tend to exhibit slower declines than the tails in normal distributions. (LO 23.e)

Module Quiz 23.3

1. **C** Correlation will be between –1 and 1. Any number above 0 is going to represent a positive output. Because rank correlation is used to measure nonlinear dependence, an output of 0.48 indicates positive nonlinear dependence. (LO 23.f)

2. **A** If all of the variables in a correlation matrix have correlations of 1, this is indicative of equicorrelation. They can have correlations of zero, as long as all are equal. Variables cannot have negative coefficients of determination (which are correlations squared) and correlations can never be greater than 1. (LO 23.g)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 13.

READING 24

SIMULATION AND BOOTSTRAPPING

Study Session 7

EXAM FOCUS

Simulation methods model uncertainty by generating random inputs that are assumed to follow an appropriate probability distribution. This reading discusses the basic steps for conducting a Monte Carlo simulation and compares this simulation method to the bootstrapping technique. For the exam, be able to explain ways to reduce Monte Carlo sampling error, including the use of antithetic and control variates. Also, understand the pseudo-random number generation method. Finally, be able to describe the advantages and disadvantages of the bootstrapping technique in comparison to the traditional Monte Carlo approach.

MODULE 24.1: MONTE CARLO SIMULATION AND SAMPLING ERROR REDUCTION

LO 24.a: Describe the basic steps to conduct a Monte Carlo simulation.

Monte Carlo simulations are often used to model complex problems or to estimate variables when the sample size is small. A few practical finance applications of Monte Carlo simulations include pricing exotic options, estimating the impact to financial markets of changes in macroeconomic variables, and examining capital requirements under stress-test scenarios.

There are five basic steps used to conduct a simulation:

1. Generate random draw data $x_i = [x_{1i}, x_{2i}, ..., x_{ni}]$. For a Monte Carlo process, this data is drawn from an assumed **data generating process (DGP)**.
2. Calculate the statistic or function of interest, $g_i = g(x_i)$.
3. Repeat steps 1 and 2 to produce N replications.
4. Estimate the quantity of interest from $\{g_1, g_2, ..., g_b\}$.
5. Evaluate the accuracy by computing the standard error. *N* should be increased until the required level of accuracy is achieved.

The first step of conducting a simulation requires generating random inputs that are assumed to follow a specific probability distribution.

The second step of the simulation generates scenarios or trials based on randomly generated inputs drawn from a pre-specified probability distribution. The most common probability distribution used is the standard normal distribution. However, Student's *t*-distribution is often used if the user believes it is a better fit for the data. A well-defined simulation model requires the generation of variables that follow appropriate probability distributions.

The third and fourth steps in the simulation process allow for data analysis related to the properties of the probability distributions of the output variables. In other words, rather than making just one output estimate for a problem, the model generates a probability distribution of estimates. This provides the user with a better understanding of the range of possible outcomes.

In step five, the quantity *N* is the number of replications or iterations and is typically performed 1,000–10,000 times, depending on how costly it is to generate the sample size.

For example, suppose we are managing an investment portfolio and desire to estimate the ending capital in the portfolio in one year, C_1. The initial capital investment, C_0, is $100 invested in the S&P 500. The return is a random variable that depends on how the market performs over the next year.

If we assume the return over the next year is equal to a historical mean return, we can calculate one point estimate of the ending capital based on the equation: $C_1 = C_0(1 + r)$. The return over the next period is a random variable, and a simulation model estimates multiple scenarios to represent future returns based on a probability distribution of possible outcomes. The output variable is an estimate of an ending amount of capital that is also a random variable. The simulation model allows us to visualize the output and analyze the probability distribution of the ending capital amounts generated by the model.

Reducing Monte Carlo Sampling Error

LO 24.b: Describe ways to reduce Monte Carlo sampling error.

The **sampling error** for a Monte Carlo simulation is quantified as the standard error estimate. The standard error of the true expected value is computed as $s/\sqrt{N}$ where *s* is the standard deviation of the output variables and *N* is the number of scenarios or replications in the simulation. Based on this equation, it intuitively follows that to reduce the standard error estimate by a factor of 10, the analyst must increase *N* by a factor of 100. (Because the square root of 100 is 10, if we increase the sample size 100 times, it will reduce the standard error estimate by dividing by 10.)

Suppose we continue the illustration from the previous example and run a simulation to estimate the ending capital amount for an initial investment portfolio of $100. The number of replications is initially 100 (i.e., N = 100), resulting in a mean ending capital

of $110 and a standard deviation of $14.80. For this example, the standard error estimate is computed as $1.48 (i.e., $14.80 / 10). Now, suppose we want to increase the accuracy by reducing the standard error estimate. How can we increase the accuracy of the simulation?

The accuracy of simulations depends on the standard deviation and the number of scenarios run. We cannot control the standard deviation, but we can control the number of replications. Assume we rerun the previous simulation with 400 replications that results in the same mean ending capital of $110, and the standard deviation remains at $14.80. The standard error estimate for the simulation with 400 replications is then $0.74 (i.e., 14.80 / 20). With four times the number of scenarios (4 × N, or 400, in this example) the standard error estimate is cut in half to $0.74. In other words, quadrupling the number of scenarios will improve the accuracy twofold.

However, increasing the number of generated scenarios can become costly for more complex multi-period simulations. Variance reduction techniques offer an alternative way to reduce the sampling error of a Monte Carlo simulation. The two most commonly used techniques for reducing the standard error estimate are antithetic variates and control variates.

Antithetic Variates

LO 24.c: Explain the use of antithetic and control variates in reducing Monte Carlo sampling error.

One reason sampling error occurs is because there are often a wide range of possible outcomes for a particular experiment or problem. Thus, to replicate the entire range of possible outcomes, the sampling sets must be recreated numerous times. However, increasing the number of samples drawn may be costly and time consuming. As an alternative approach, the **antithetic variate technique** can reduce Monte Carlo sampling error by rerunning the simulation using a *complement* set of the original set of random variables.

If the original set of random draws is denoted u_t for each replication, then the simulation is rerun with the complement set of random numbers denoted $-u_t$. The use of antithetic variates should result in a lower covariance and variance, because the two sets are perfectly negatively correlated [i.e., $\text{corr}(u_t, -u_t) = -1$]. The following example illustrates how the standard error for a Monte Carlo simulation is reduced by using the antithetic variate technique.

First, consider a simulation of two sets that does not use the antithetic variate technique. Suppose the average parameter estimate is determined by two Monte Carlo simulations using different random sample sets. The average output parameter value $\bar{x}$ for the two simulations using different random sample replications is simply calculated as:

$$\bar{x} = (x_1 + x_2) / 2$$

where x_1 and x_2 are the average output parameter values for simulation sets one and two, respectively.

Next, we can calculate the variance of the average of the two sets as follows:

$$\text{var}(\bar{x}) = \frac{\text{var}(x_1) + \text{var}(x_2) + 2\text{cov}(x_1, x_2)}{4}$$

Without using antithetic variates, the two sets of Monte Carlo replications are independent. Thus, the covariance will be zero and the variance of $\bar{x}$ is simply reduced to the following:

$$\text{var}(\bar{x}) = \frac{\text{var}(x_1) + \text{var}(x_2)}{4}$$

The use of antithetic variates results in negative covariance between the original random draws and their complements (i.e., antithetic variates). This negative relationship means that the Monte Carlo sampling error must always be smaller using this approach.

Control Variates

The **control variate technique** is a widely used method to reduce the sampling error in Monte Carlo simulations. A control variate involves replacing a variable *x* (under simulation) that has unknown properties, with a similar variable *y* that has known properties.

Suppose two separate simulations are conducted on variable *x* with unknown properties and control variable *y* with known properties using the same set of random numbers. Also assume that the Monte Carlo simulation estimated variables for x and y are denoted as $\hat{x}$ and $\hat{y}$, respectively. The original estimate for *x* can be redefined as *x** as follows:

$$x^* = y + (\hat{x} - \hat{y})$$

The new *x** variable estimate will have a smaller sampling error than the original *x* variable if the control statistic and statistic of interest are highly correlated. The Monte Carlo results for the new *x** variable are assumed to have similar properties to the known *y* control variable.

The following mathematical equations help illustrate the condition that is necessary to reduce the sampling error using control variates. Consider taking the variance of both sides of the equation that defines the new variable such that:

$$\text{var}(x^*) = \text{var}\left[y + (\hat{x} - \hat{y})\right]$$

The control variable *y* does not have a sampling error because it has known properties. Thus, *var(y)* equals zero. Now, the variance of the remaining two variables can be rewritten as follows:

$$\text{var}(x^*) = \text{var}(\hat{x}) + \text{var}(\hat{y}) - 2\text{cov}(\hat{x}, \hat{y})$$

The control variate method will only reduce the sampling error in Monte Carlo simulations if *var(x*)* is less than $\hat{x}$. Another way of expressing this condition is as

follows:

$$\text{var}(\hat{y}) - 2\text{cov}(\hat{x},\hat{y}) < 0$$

This relationship can be simplified as follows:

$$\text{cov}(\hat{x},\hat{y}) > \frac{\text{var}(\hat{y})}{2}$$

The covariance can be converted to correlation by dividing both sides of the previous inequality by the product of the standard deviations as follows:

$$\text{corr}(\hat{x},\hat{y}) > \frac{1}{2}\sqrt{\frac{\text{var}(\hat{y})}{\text{var}(\hat{x})}}$$

A practical financial example of applying control variates is the use of Monte Carlo simulations in pricing Asian options. An Asian option is priced based on the average value of the underlying asset over the lifespan of the option. The use of a similar derivative, such as a European option with known statistical properties, can be used as a control variate. The price of the European option PBS is determined by the Black-Scholes-Merton option pricing model. Next, simulated prices are determined for the Asian option and the European option and denoted P_A and P_{BS}*, respectively. The new estimate of the Asian option price P_A* can then be determined based on the following equation:

$$P_A^* = (P_A - P_{BS}) + P_{BS}^*$$

MODULE QUIZ 24.1

1. Which of the following statements regarding Monte Carlo simulation is least accurate? When using Monte Carlo simulation:
   A. simulated data is used to numerically approximate the expected value of a function.
   B. the user specifies a complete data generating process (DGP) that is used to produce simulated data.
   C. the observed data are used directly to generate a simulated data set.
   D. a full statistical model is used that includes an assumption about the distribution of the shocks.

2. Suppose an analyst is concerned about Monte Carlo sampling error. Based on an initial Monte Carlo simulation with 100 replications, the results indicated a standard deviation of 12.64. The simulation was rerun with 900 replications and the standard deviation remained at 12.64. What are the standard error estimates for the simulations with 100 replications and 900 replications, respectively?

	N = 100	N = 900
A.	0.126	0.014
B.	0.126	0.140
C.	1.264	0.421
D.	1.264	0.214

3. A concern for Monte Carlo simulations is the size of the sampling error. One way to reduce the sampling error is to use the antithetic variate technique. Which of the following statements best describes this technique?
   A. The simulation is rerun using a complement set of the original set of random variables.
   B. The number of replications is increased significantly to reduce sampling error.

C. Sample data is replaced after every replication to ensure it has an equal probability of being redrawn.
D. The data generating process (DGP) is approximated by redefining the unknown variable with a variable that has known properties.

MODULE 24.2: BOOTSTRAPPING AND RANDOM NUMBER GENERATION

The Bootstrapping Method

LO 24.d: Describe the bootstrapping method and its advantage over Monte Carlo simulation.

Another way to generate random numbers is the **bootstrapping method**. The bootstrapping approach draws random return data from a sample of historical data. Unlike the Monte Carlo simulation method, bootstrapping uses actual historical data instead of random data from a probability distribution. Furthermore, bootstrapping repeatedly draws data from the historical data set and replaces the data so it can be drawn again.

Unlike Monte Carlo simulation, bootstrapping does not directly model the observed data, nor does it make assumptions about the distribution of the data. Rather, the observed data is sampled directly from the unknown distribution.

There are two commonly used classes of bootstraps: independent and identically distributed (i.i.d.) and circular block bootstrap (CBB).

Independent and identically distributed (i.i.d.)

The first bootstrapping approach that we will consider is the i.i.d. bootstrap. In this methodology, samples are simply drawn one-by-one from the observed data, and replaced.

If we require a simulation of sample size of three from a data set with a total of 10 observations, the i.i.d. bootstrap generates observation indices by randomly sampling three times with replacement from the values {1, 2, … , 10}. These indices indicate the observed data to be included in the simulated (i.e., bootstrap) sample.

For example, suppose that the three observations are drawn from a sample of 10 data points $\{x_1, x_2 \dots, x_{10}\}$. The first simulation might include observations $\{x_2, x_7, x_9\}$, the second simulation $\{x_2, x_5, x_{10}\}$, and the third $\{x_1, x_1, x_8\}$. Note that the first two simulated samples overlap (both contain x_2), which is possible because the i.i.d. bootstrapping method samples *with* replacement. Notice also that the third simulation sample includes the same observation (x_1) twice. This too is a result of sampling with replacement.

The i.i.d. bootstrap methodology is valid when the observations are independent. However, in finance it is often the case that data is *dependent* across time; for example, volatility tends to be high during some periods and low during other periods.

Circular block bootstrap (CBB)

When observations are not independent, a more sophisticated bootstrapping method than i.i.d. is required. One such method is the CBB method. CBB differs from i.i.d. in that rather than sampling single observations, the CBB method samples *blocks* of observations. The CBB method is used to produce bootstrap samples by sampling blocks, with replacement, until the required bootstrap sample size is produced.

For example, suppose that 10 observations are available and they are sampled in blocks of size three. Ten blocks are constructed, starting with $\{x_1, x_2, x_3\}$, $\{x_2, x_3, x_4\}$, ...,$\{x_8, x_9, x_{10}\}$, $\{x_9, x_{10}, x_1\}$, $\{x_{10}, x_1, x_2\}$. Notice that the first eight blocks use three consecutive observations, but the final two blocks *wrap around*.

The block size used in the CBB methodology should be large enough to reflect the dependence in the data. However, the block size should not be so large that the number of blocks becomes small. For a sample size of *n*, a block size of $\sqrt{n}$ is generally appropriate.

LO 24.f: Describe situations where the bootstrapping method is ineffective.

While bootstrapping is a useful statistical technique, it has its limitations. There are two specific issues that arise when using a bootstrap:

1. *Using the entire data set may not be reliable*: As long as current market conditions are normal, using the complete historical data set is beneficial. On the other hand, when the current condition of financial markets is different from its usual state, the bootstrapping method may be ineffective. For example, using a bootstrap to estimate the value at risk during the financial crisis of 2007–2009 would produce an unrealistically low view of risk, because volatility was historically much lower.
2. *Structural changes*: Another limitation of the bootstrapping method is that there may have been recent permanent fundamental changes in the market. For example, interest rates on U.S. T-bills were near-zero for a decade beginning in 2008—a condition that had never previously occurred over a long period. As a result, bootstrapping using older historical data would be ineffective in replicating this period.

Random Number Generation

LO 24.e: Describe pseudo-random number generation.

Random number generators are used to produce an irregular sequence of numerical values. Algorithms used to generate these random sequences are referred to as **pseudo-random number generators (PRNGs)**. The term *pseudo* implies that these computer-generated numbers are *not truly random*: they are actually generated from a formula.

PRNGs typically produce sequences of random numbers uniformly distributed between zero and one. Each number should have an equal probability of being drawn from the

uniform (0,1) distribution.

To produce pseudo-random numbers, an initial seed value must first be chosen. The choice of seed value will determine the random number sequence that is generated. In fact, any particular seed value will generate an identical set of values each time the PRNG is run.

The recurring nature of PRNG outputs provides us with two benefits:

1. *Repeatability*: Because a particular seed value will always produce the same series of random values, we can replicate the sequence across several different experiments, which allows multiple alternative models to be estimated using the same simulated data. Furthermore, the use of a specific initial seed allows simulation results to be reproduced later—which may be required for regulatory compliance.
2. *Computing Clusters*: Suppose that we are using a group of computers to model complex portfolios containing thousands of financial instruments that are all impacted by the same set of fundamental factors. Using a common seed value allows us to use the same set of random numbers across multiple simulations. Starting each PRNG in a cluster with the same seed allows each simulation to make use of the same values when studying the joint behavior of the instruments in the portfolio.

Disadvantages of Simulation Approaches

LO 24.g: Describe the disadvantages of the simulation approach to financial problem solving.

Disadvantages of the simulation approach to financial problem solving include:

1. *Specification of the DGP*: Even with a large number of simulation iterations, when the assumptions of model inputs or the data generating process are unrealistic, imprecise results may occur. Alternate assumptions made in the DGP may lead to substantially different results. A common model misspecification relates to assumptions about the underlying probability distribution of inputs: for example, option prices are typically fat-tailed, but a model could erroneously draw option prices from a normal distribution. This would lead to inaccurate results, regardless of the number of replications.
2. *Computational cost*: The best way to reduce the variation of simulation results is to use a large number of replications. If estimated parameters are complex, the computations may take an extremely long time to run. Some problems may require a large number of replications to obtain acceptable results; it is common to use at least 10,000 replications in Monte Carlo simulations. Computer processor times have improved exponentially, however. The complexity of markets and issues that are examined have also become increasingly complex, potentially leading to *high computation costs*.

MODULE QUIZ 24.2

1. Which of the following statements regarding the bootstrapping method is least accurate? Bootstrapping simulations:

A. draw data from historical data sets.
B. replace drawn data so it can be redrawn.
C. require assumptions with respect to the true distribution of the parameter estimates.
D. rely on the key assumption that the present resembles the past.

2. Which of the following statements regarding the pseudo-random number generation method is least accurate? Pseudo-random numbers are:
   A. not truly random.
   B. actually generated from a formula.
   C. determined by the choice of the initial seed value.
   D. impossible to predict.

3. The bootstrapping method is most likely to be effective when the:
   A. data contains outliers.
   B. present is different from the past.
   C. data is independent.
   D. markets have experienced structural changes.

4. Monte Carlo simulation is a widely used technique in solving economic and financial problems. Which of the following statements is least likely to represent a limitation of the Monte Carlo technique when solving problems of this nature?
   A. High computational costs arise with complex problems.
   B. Simulation results are experiment-specific because financial problems are analyzed based on a specific data generating process (DGP) and set of equations.
   C. Results of most Monte Carlo experiments are difficult to replicate.
   D. If the input variables have fat tails, Monte Carlo simulation is not relevant because it always draws random variables from a normally distributed population.

KEY CONCEPTS

LO 24.a

A Monte Carlo simulation uses observations to estimate key model parameters, such as the mean and standard deviation. A complete data generating process (DGP) is created by combining these parameters with an assumption about the distribution of the standardized returns.

The basic steps of a Monte Carlo simulation are:

1. Generate data according to the assumed DGP.
2. Calculate the function or statistic of interest.
3. Repeat steps one and two to produce N replications.
4. Estimate the quantity of interest.
5. Assess the accuracy by computing the standard error, and increase N until the required accuracy is achieved.

LO 24.b

The standard error estimate of a Monte Carlo simulation, $s/\sqrt{N}$, can be reduced by a factor of 10 by increasing *N* by a factor of 100, where quantity *N* is the number of replications or iterations.

LO 24.c

Antithetic variables and control variates can be used simultaneously to reduce the approximation error in a Monte Carlo simulation.

With antithetic variables, random values are constructed to generate negative correlation within the values used in the simulation. Variance is reduced because the covariance between the simulated values is negative, so the variance of the sum is less than the sum of the variances.

Control variates reduce the variance of the approximation by adding values with a mean of zero that are correlated to the simulation.

LO 24.d

Bootstrapping simulations repeatedly draw data from historical data sets, each time replacing the data so it can be redrawn. The bootstrapping technique requires no assumptions with respect to the true distribution of the parameter estimates.

LO 24.e

Pseudo-random numbers are not truly random, as they are actually generated from a formula. The choice of the initial seed value determines the random numbers that are generated.

The reproducibility of outputs from pseudo-random number generators (PRNGs) allows results to be replicated across multiple experiments, or to be generated on multiple computers.

LO 24.f

Two primary limitations arise when using the bootstrapping method:

1. This method may not be reliable if current financial market conditions differ from their normal state.
2. Structural changes may have occurred in markets so that current conditions are different from anything that has occurred in the past.

LO 24.g

Two disadvantages of the simulation approach to financial problem solving include:

1. Specification of the data generating process (DGP): When the assumptions of model inputs or the data generating process are unrealistic, inaccurate results may occur.
2. Computational cost: While computer processing times have decreased, markets have also become increasingly complex, potentially leading to high computation costs.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 24.1

1. **C** In both Monte Carlo simulation and bootstrapping, the goal is to numerically approximate the expected value of a complex function through the use of computer-generated values (i.e., simulated data). The main difference between Monte Carlo simulation and bootstrapping is the source of the simulated data: in Monte Carlo simulation, the user specifies a complete DGP that is used to produce

the simulated data, while in bootstrapping, the observed data are used directly to generate the simulated data set—without specifying a complete DGP. (LO 24.a)

2. **C** The standard error is determined by dividing the standard deviation by the square root of the number of replications $s/\sqrt{N}$. The standard error estimate for the first simulation of 100 replications is 1.264 (i.e., 12.64 / 10). With 900 replications, the standard error estimate is reduced to 0.4213 (i.e., 12.64 / 30). (LO 24.b)
3. **A** The antithetic variate technique reduces Monte Carlo sampling error by rerunning the simulation using a complement set of the original set of random variables. (LO 24.c)

Module Quiz 24.2

1. **C** The bootstrapping technique does not require any assumptions with respect to the true distribution of the parameter estimates. Bootstrapping simulations repeatedly draw data from historical data sets, and then replace the data so it can be redrawn. The bootstrapping method is only as valid as the assumption that the present resembles the past. (LO 24.d)
2. **D** Pseudo-random numbers appear random because they are *difficult* to predict. However, they are produced by deterministic functions that are complex rather than truly random. The initial choice of a seed value determines the series of random numbers that is generated. (LO 24.e)
3. **C** The bootstrapping method is most likely to be effective when the data is independent and there are no outliers in the data. Bootstrapping uses the entire data set to generate a simulated sample, so the bootstrapping method should be reliable if the current state of the financial market is the same as its normal state, meaning that no structural changes have taken place. (LO 24.f)
4. **D** A disadvantage of Monte Carlo simulations is that imprecise results may occur when the assumptions of model inputs or DGP are unrealistic. The distribution of input variables does not need to be the normal distribution. Problems will arise if a real-world variable is fat-tailed, but the model erroneously draws option prices from a normal distribution. (LO 24.g)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 14.

READING 25

MACHINE LEARNING MODELS

Study Session 7

EXAM FOCUS

The focus of this reading is on machine learning models and how they can be applied in financial and operational situations. For the exam, understand the differences between traditional econometric models and machine learning approaches, as well as the different categories of machine learning. Also, be familiar with principal components analysis (PCA) and K-means algorithms, which are concepts used to reduce dimensionality and create data clusters, respectively. Modeling challenges include the risks of underfitting and overfitting data, as well as differentiating between the different data sub-samples (i.e., training set, validation set, test set). Finally, additional concepts in this reading include reinforcement learning, which is an established set of decisions designed to maximize reward, and natural language processing (NLP), which uses algorithms to interpret written and verbal human language.

MODULE 25.1: MACHINE LEARNING AND DATA PREPARATION

Machine Learning vs. Classical Econometrics

LO 25.a: Discuss the philosophical and practical differences between machine-learning techniques and classical econometrics.

Machine learning represents a range of techniques where models recognize data patterns for practical applications. Compared to traditional statistical techniques, machine learning is designed to handle extremely large volumes of data (i.e., big data), provide greater flexibility, and employ a wide range of specifications (e.g., capturing non-linear interactions which standard linear models cannot).

Under classical statistics and econometrics analyses, economic and/or financial theory drives the data-generating process. An analyst chooses the model and variables, while a computer algorithm is used to estimate parameters and test their significance. The analyst must use the results to determine if the data supports the expected outcome. On the other hand, machine learning allows data to decide what the models will include, with no specific hypothesis from an analyst tested as part of the process.

Machine learning functions well when financial theory does not dictate the choice of variables or when incorporating linear specifications is undetermined. Machine learning models use flexible functional forms to capture complex (and potentially non-linear) variable interactions, which differs from what is required with linear regression models. Explanatory variable independence, normal distributions, statistical significance, error-term testing, and quality of regression fit are critical for conventional models, but they are not applicable to "supervised" machine learning, which is focused on prediction accuracy.

The differences in key terminology between conventional econometrics and machine learning are shown in Figure 25.1.

Figure 25.1: Terminology Differences

Conventional Econometrics	Machine Learning
Independent Variables	Inputs or Features (values)
Dependent Variables	Outputs or Targets (forecasted values, known as labels)

Supervised, Unsupervised, and Reinforcement Learning Models

LO 25.g: Differentiate among unsupervised, supervised, and reinforcement learning models.

Machine learning methodologies land in the following three categories:

- **Supervised learning**: used to predict either the value of a variable (e.g., the value of a car) or the classification of an observation (e.g., the outcome of a sports team's next game—win or loss). The algorithm uses "labeled data" to learn. For the value of a car, this data would include the make, model, year, size, engine type, etc. For the outcome of a sporting event, this data would include current record, location of the event, player years of experience, etc.
- **Unsupervised learning**: pattern recognition in data with no specific target. Data clustering and the identification of a small group of explanatory factors may be involved.
- **Reinforcement learning**: incorporates a trial-and-error approach to make decisions in a changing environment.

Supervised learning can be used to predict a variable's value in a time-series (predicting Treasury note yields one year from now) or a cross-sectional scenario for data not in a sample (the list price for a house based on comparable prices in the neighborhood).

Unsupervised learning is not used for predictions; however, it is very useful for characterizing and learning the structure of a dataset. An internal auditor may use unsupervised learning to evaluate various company transactions on specific ledger accounts to determine if further review and investigation is needed.

Reinforcement learning is often used in risk management contexts such as large block trading, hedging derivatives portfolios, or managing overall investment portfolios.

Data Preparation and Cleaning

For **data preparation**, the two primary methods used to achieve scale consistency within the dataset, which is often required by machine learning models, are standardization and normalization.

- **Standardization**: this is a process used to create a scale for measuring variables with zero mean and unit variance. This is the preferred methodology for data covering a wide scope (including outliers). The process involves subtracting the sample mean of each variable from all observations and dividing by the standard deviation.

  $$\tilde{x}_{i,j} = \frac{x_{i,j} - \hat{\mu}_i}{\hat{\sigma}_i}$$

  where:

  $\hat{\mu}_i$ = estimated mean

  $\hat{\sigma}_i$ = estimated standard deviation

- **Normalization**: this process, also called min-max transformation, creates a variable between zero and one which will not usually have a zero mean or unit variance.

  $$\tilde{x}_{i,j} = \frac{x_{i,j} - x_{i,min}}{x_{i,max} - x_{i,min}}$$

  where:

  $x_{i,min}$ = minimum of the observations

  $x_{i,max}$ = maximum of the observations

Data cleaning is a critical and time-consuming component of machine learning approaches. Several reasons for data cleaning include:

- Missing data. Although this is the most common problem, it can be remedied by either removing observations with missing data (if it's a small number), replacing the missing observations with the mean or median of observations on the feature, or estimating them based on other observations.
- Outliers (i.e., observations several standard deviations away from the mean).
- Duplicate observations.
- Inconsistent recording.
- Unwanted (i.e., irrelevant) observations.

MODULE QUIZ 25.1

1. Compared to traditional statistical methodologies, machine learning provides all of the following benefits except:
   A. greater flexibility.
   B. there is no need to scale the data.
   C. the ability to manage large volumes of data.
   D. the capacity to capture non-linear transactions.

2. The compliance manager at a bank uses machine learning approaches to review journal entries posted to the bank's general ledger. In particular, she is concerned with employees using the wrong ledger accounts to record transactions. This type of machine learning is best categorized as:
   A. supervised learning.
   B. unsupervised learning.
   C. reinforcement learning.
   D. linear regression analysis.

MODULE 25.2: PRINCIPAL COMPONENTS ANALYSIS AND K-MEANS CLUSTERING

Principal Components Analysis

LO 25.d: Use principal components analysis to reduce the dimensionality of a set of features.

A popular statistical technique for dimension reduction in unsupervised learning models is **principal components analysis (PCA)**. The goal of PCA is to produce almost the same amount of information using a small number of uncorrelated components (i.e., variables) that a large number of correlated components can provide. Thus, in a machine learning model, PCA is used to reduce the number of features.

PCA is often applied to yield curve movements by producing a small count of uncorrelated components that describe the movements of the curve. The observed components should represent a linear combination of the variables used. The two most important explanatory components are the parallel shift (all rates move in the same direction by the same amount) and the twist (where short-term and long-term rates move in opposite directions).

Based on a review of seven Treasury rates over a 10-year period (120 months), the first three observed components were responsible for 99% of the overall variation in yield movements due to the high correlation between yield movements.

K-Means Clustering

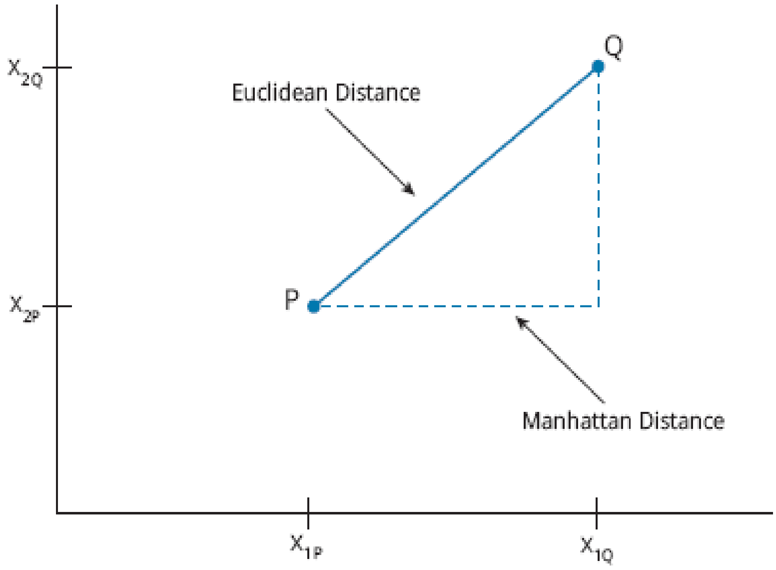
LO 25.e: Describe how the K-means algorithm separates a sample into clusters.

To identify the structure of a dataset, an unsupervised **K-means algorithm** can be used to separate dataset observations into clusters. The value *K* represents the number of clusters and is set by an analyst. The centers of the data clusters are called **centroids** and are initially randomly chosen. Each data point is allocated to its nearest centroid and then the centroid is recalculated to be at the center of all the data points assigned to it. This process continues until the centroids remain constant.

The distance of each data point to its centroid can be calculated using either the Euclidean or the Manhattan distance measure. Assume there are two features (x_1 and

x_2) and two observations (P and Q) as shown in Figure 25.2.

Figure 25.2: Euclidean and Manhattan Distances



Euclidean distance is the diagonal line between P and Q (i.e., the triangle hypotenuse) in Figure 25.2. It can be computed using two features or *m* features as follows:

$$\text{Two features: } d_E = \sqrt{(x_{1Q} - x_{1P})^2 + (x_{2Q} - x_{2P})^2}$$

$$m \text{ features: } d_E = \sqrt{\sum_{i=1}^{m}(x_{iQ} - x_{iP})^2}$$

Manhattan distance approximates the distance between P and Q via the path of the opposite and adjacent sides of Figure 25.2.

$$\text{Two features: } d_M = |x_{1Q} - x_{1P}| + |x_{2Q} - x_{2P}|$$

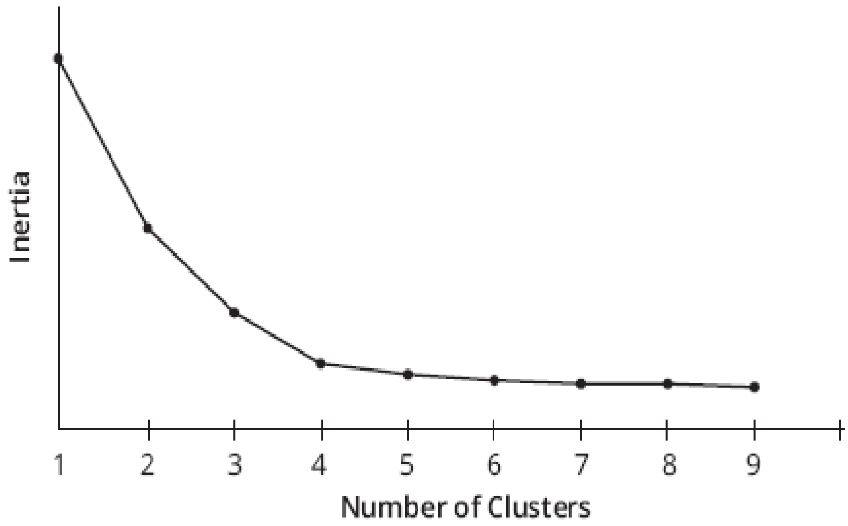
$$m \text{ features: } d_M = \sum_{i=1}^{m}|x_{iQ} - x_{iP}|$$

The goal of the K-means algorithm is to minimize the distance between each observed point and its centroid. A model's fit is better when the individual data points are close to their centroid. **Inertia**, a measure of the distance (*d*) between each data point (*j*) and its centroid, is defined as:

$$\text{inertia} = \sum_{j=1}^{n} d_j^2$$

A lower inertia implies a better cluster fit. However, because inertia will always fall as more centroids are added, there is a limit to which adding more centroids adds value. A "scree plot" may be employed to calculate inertia for different values of *K*. The plot is used to evaluate where inertia starts to decline at a slower pace as *K* increases. This "elbow" of the plot represents the optimal number of centroids (according to Figure 25.3, that value would be 4).

Figure 25.3: Determining the Number of Clusters (Scree Plot)



As an alternative approach, a **silhouette coefficient** can be used to choose *K* by comparing the distance between an observation and other points in its own cluster to its distance to data points in the next closest cluster. The highest silhouette score will produce the optimal value of *K*.

MODULE QUIZ 25.2

1. The end goal of principal components analysis (PCA) is the use of which of the following to manage dimensionality?
   A. A small number of correlated components.
   B. A large number of correlated components.
   C. A small number of uncorrelated components.
   D. A large number of uncorrelated components.
2. The optimal number of centroids can be found by choosing the:
   A. value that produces the lowest possible inertia.
   B. value that produces the highest possible inertia.
   C. point where inertia declines at a faster pace as K increases.
   D. point where inertia declines at a slower pace as K increases.

MODULE 25.3: METHODS OF PREDICTION AND SAMPLE SPLITTING

Underfitting and Overfitting

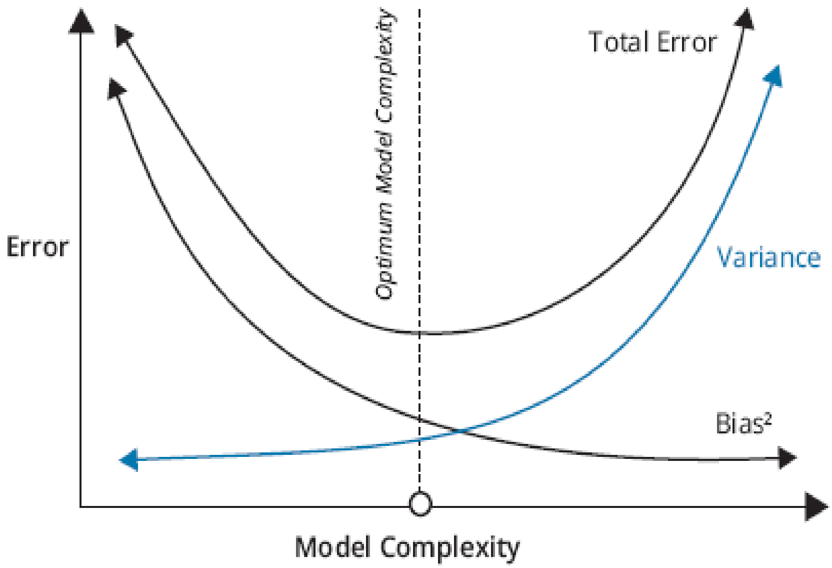
LO 25.c: Understand the differences between and consequences of underfitting and overfitting, and propose potential remedies for each.

Overfitting occurs when a model is too complex, too large, or has too many parameters; this is a greater risk for machine learning models than for conventional econometric models (which typically have limited parameters). Overfitted models are evident when new data causes the model to perform worse. Model construction incorporates a training data set (used to estimate model parameters) and a validation data set (used to evaluate performance on a separate data set). The output of an overfitted model is a training set with a very low error rate, but poor performance when applying the model to data outside of the training set.

Underfitting occurs when a model is too simple and fails to capture relevant patterns. This is illustrated by the expected performance of hedge funds relative to the size (i.e., assets under management) of their operations. An expected outcome would be quadratic, with small funds struggling to cover costs and larger funds struggling to implement timely strategic decisions. A machine learning model (with no assumptions regarding model structure) would appropriately address this non-linear relationship, while a conventional model would underfit the data. As such, there is a greater risk of underfitting in conventional models.

The size of the machine learning model will determine whether the model is appropriately fitted, overfitted, or underfitted. An underfitted model excludes relevant factors, which will result in biased predictions with a low variance. An overfitted model has the opposite result, with low bias but high variance predictions. The **bias-variance tradeoff** is illustrated in Figure 25.4.

Figure 25.4: Bias-Variance Tradeoff



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Training, Validation, and Test Data Sub-Samples

LO 25.b: Explain the differences among the training, validation, and test data sub-samples, and how each is used.

A common modeling technique in both conventional econometric models and machine learning models is keeping part of the data sample out of the model (i.e., holdout data) to test the fitted model. The goal is to see how well the retained part of the data sample can predict unseen observations. With machine learning models, the data sample is often divided into three sections: training, validation, and testing.

- **Training set**: employed to estimate the parameters of the model. This set represents the data used by the machine learning model to learn how to best represent the data.

- **Validation set**: used to decide between two alternative models. This data can no longer be used once the superior model has been chosen.
- **Test set**: used to determine the effectiveness of the chosen model. A model is effective when the fit to the test sample is almost as good as the fit to the training sample (known as model *generalization*).

Although there is no set allocation for how much a given sample should go to the respective sets above, a typical allocation is two-thirds of the data going to the training set, one-sixth going to the validation set, and the other one-sixth going to the test set. Biases in parameter estimation are a risk associated with a small training set, while inaccurate model estimations are a risk of a small validation set.

Cross-sectional output data has no natural order, which allows for random placement of the data into the three sets described. Time-series data has a natural order, so training data will typically be the first part of the sample, followed by validation data, and then test data.

The larger the data set, the lower the risk of improper allocations. If the data set is relatively small, *k*-fold **cross-validation** may be utilized. This technique combines training and validation data into a single sample, with the combined data (*n*) allocated into *k* samples. A common choice is $k = 5$ or 10, so if $k = 10$, then the data will be divided into 10 randomly selected sub-samples of equivalent size (i.e., 10% of the data). The first estimation would use k1 through k9, with the k10 sample removed (and reserved for the test set). The next would use k1 through k8 and k10, with k9 removed. This creates 10 validation samples that are averaged to assess performance. A small number of observation points may benefit from a larger *k* and setting $k = n$ is known as *leave-one-out cross-validation*, where the number of folds is equal to the number of data points in the set.

MODULE QUIZ 25.3

1. The predictions that are generated from an underfitted model will likely have:
   A. low bias and low variance.
   B. low bias and high variance.
   C. high bias and low variance.
   D. high bias and high variance.
2. An analyst is choosing between two machine learning models. Which of the following datasets will the analyst most likely use to make the determination of which model to select?
   A. Test set.
   B. Training set.
   C. Variance set.
   D. Validation set.

MODULE 25.4: REINFORCEMENT LEARNING AND NATURAL LANGUAGE PROCESSING

Reinforcement Learning

LO 25.h: Explain how reinforcement learning operates and how it is used in decision-making.

Reinforcement learning involves the creation of a policy for decision-making, with the goal of maximizing reward. Much like how computers learn to play games, a systematic approach involving trial and error is incorporated by having the computer play against itself until it learns which steps to follow in every possible situation. Applications are seen in finance in areas such as derivatives hedging, trading rules, and managing large-volume trades. Reinforcement learning algorithms require large quantities of training data, and initial performance will be weak, but will improve dramatically over time.

The key areas of reinforcement learning are known as states, actions, and rewards:

- **States (S)**: define the environment.
- **Actions (A)**: represent the decisions taken.
- **Rewards (R)**: maximized when the best possible decision is made.

The *Q-value* is the expected value of taking an action (*A*) in a certain state (*S*). The best action to take in any given state (*S*) is whatever the value of *A* is that maximizes the expression below:

$$V(S) = \max_A(Q(S, A))$$

To determine actions taken for each state, the algorithm will choose between the best action already identified (known as **exploitation**) and a new action (known as **exploration**). The probability assigned to exploitation and exploration is p and $1 - p$, respectively. As more trials are completed and the algorithm has learned the superior strategies, the value of p increases.

The **Monte Carlo method** may be deployed to evaluate actions (*A*) taken in states (*S*) and the subsequent rewards (*R*) that may result. The formula is shown as follows, with the α parameter set at a number like 0.01 or 0.05.

$$Q^{new}(S, A) = Q^{old}(S, A) + \alpha\left[R - Q^{old}(S, A)\right]$$

The **temporal difference learning method**, an alternative to the Monte Carlo method, assumes the best strategy thus far is the one to be made going forward and will only look one decision ahead.

An example of reinforcement learning is shown using the data in Figure 25.5, which represents the current Q(S, A) values. If Action 3 in State 4 is taken as the next trial, assume the subsequent reward is 1.2. If α is 0.01, the Monte Carlo method would result in Q(4, 3) being updated from 0.5 to 0.5 + 0.01(1.2 – 0.5) = 0.507.

If the next decision to be made is in State 3, assume the reward earned between the two decisions is 0.3. The value of being in State 3 (Action 3) is 0.8. If α is 0.01, the temporal difference method will result in Q(4, 3) adjusted from 0.5 to 0.5 + 0.01(0.3 + 0.8 – 0.5) = 0.506.

Figure 25.5: Current Q-Values

	State 1	State 2	State 3	State 4
Action 1	0.3	0.6	0.1	0.4
Action 2	0.7	0.3	0.4	0.2
Action 3	0.9	0.6	0.8	0.5

Deep reinforcement learning occurs when **neural networks** are used to estimate a complete table from available observations.

Natural Language Processing

LO 25.f: Be aware of natural language processing and how it is used.

Natural language processing (NLP) is a component of machine learning focused on understanding and analyzing written and verbal human language. NLP, also known as **text mining**, has been employed in accounting fraud detection, assessing sentiment of corporate statements/comments, categorizing text, and recognizing specific words for determining the purpose of a message. NLP offers the benefit of speed and document review without inconsistencies or bias found in human reviews.

The steps in NLP include capturing the language in a document, preprocessing the text, and analyzing it for a specific purpose. Preprocessing text requires the following steps:

1. The document must be **tokenized**, which means identifying only the words (i.e., removing punctuation, symbols, and spacing) and modifying all of them into lowercase.
2. Removing **stopwords** such as "the," "has," and "a." These words are designed to make sentences flow but have no other value.
3. **Stemming**, which means replacing words with their stems. For example, "arguing," "argued," and "argues" maps to "augu."
4. **Lemmatization**, which is replacing words with their lemmas. For example, "worse" maps to "bad." This is a similar concept to stemming, but the lemma will be an actual word.
5. **N-grams** may be considered, which are groups of words that have meaning when placed together as opposed to being considered individually. For example, the trigram "exceed analyst expectations" may be more meaningful than the separate words "exceed," "analyst," and "expectations."

The processed text is considered by NLP as a **"bag of words"** where the order and linkages of words (outside of "N-grams") are irrelevant.

NLP is often used to assess whether things like corporate news releases are considered positive, negative, or neutral. Using an inventory of "sentiment words," counts of words pre-classified as positive, negative, or neutral are compared to evaluate the overall sentiment of a news release. Examples of positive word stems are grow, relief, increase, and rise. Examples of negative word stems are concern, disappoint, decrease, and decline.

In a situation where a news release may have four positive words and six negative words, the sentiment of the overall piece would be considered negative. A challenge is how to handle situations where positive words are found in sentences with a negative connotation (e.g., "would have resulted in an increase"). Comparisons of human reviews versus algorithm reviews of the same news release may be useful in refining the accuracy of this approach.

MODULE QUIZ 25.4

1. An analyst applying a reinforcement learning model has assigned a probability to exploitation of 65%. As she completes more trials, she can reasonably expect that the probability will increase above:
   A. 35% for exploration.
   B. 65% for exploitation.
   C. 50% for exploration.
   D. 50% for exploitation.

2. Natural language processing (NLP) is used to evaluate the MD&A (Management, Discussion, and Analysis) section of a company's annual report. In removing the "stopwords," the NLP algorithm will remove all of the following words except:
   A. "or."
   B. "are."
   C. "have."
   D. "fallen."

KEY CONCEPTS

LO 25.a

Machine learning (ML) involves using models to recognize data patterns for practical applications. Compared to traditional statistical techniques, machine learning is designed to handle extremely large volumes of data, provide greater flexibility, and employ a wider range of specifications. With machine learning, the data drives what the models will include, with no specific hypothesis from an analyst tested as part of the process.

Machine learning models use flexible functional forms to capture complex (and potentially non-linear) variable interactions, which differs from what is required with linear regression models. The focus of machine learning is on prediction accuracy. Conventional econometrics uses independent and dependent variables, whereas machine learning uses inputs/features and outputs/targets.

LO 25.b

The data sample in machine learning models is often divided into three parts: training, validation, and testing. The training set represents the data used by the machine learning model to learn how to best represent the data. The validation set is used to decide between two alternative machine learning models. The test set is used to determine the effectiveness of the chosen model.

A typical allocation is two-thirds of the data going to the training set, one-sixth going to the validation set, and the other one-sixth going to the test set. Biases in parameter estimation are a risk associated with a small training set, while inaccurate model estimations are a risk of a small validation set.

Cross-sectional output data has no natural order, which allows for random placement of the data into the three sets described above. Time-series data has a natural order, so training data will typically be the first part of the sample, followed by validation data and then test data.

If the data set is relatively small, k-fold cross-validation may be utilized where training and validation data are combined (n) and allocated into k samples. Setting $k = n$ is known as leave-one-out cross-validation where the number of folds is equal to the number of data points in the set.

LO 25.c

Overfitting, which is a greater risk for machine learning models, occurs when a model is too complex, too large, or has too many parameters. Overfitted models are evident when new data causes the model to perform worse. Underfitting occurs when a model is too simple and fails to capture relevant patterns. There is a greater risk of underfitting in conventional models.

The size of the machine learning model will determine whether the model is appropriately fitted, overfitted, or underfitted. As mentioned, an underfitted model excludes relevant factors, which will result in biased predictions with a low variance. An overfitted model has the opposite result, with low bias but high variance in predictions among datasets.

LO 25.d

Principal components analysis (PCA) is used for dimension reduction, with a goal of producing almost the same amount of information using a smaller number of uncorrelated components (variables) than a large number of correlated components can provide.

PCA is often applied to yield curve movements. The two most important explanatory components are the parallel shift (all rates move in the same direction by the same amount) and the twist (where short-term and long-term rates move in opposite directions). It is often the case that the top three components are responsible for almost all the variation in yield movements.

LO 25.e

The K-means algorithm can be used to separate observations into clusters, where K represents the number of clusters and is set by an analyst. The centers of the clusters are called centroids.

The distance of each data point to its centroid can be calculated using either the Euclidean distance or the Manhattan distance. The goal of the K-means algorithm is to minimize the distance between each observed point and its cluster centroid. A model's fit is better when the individual data points are close to their centroid.

Inertia is a measure of the distance between each data point and its centroid. A lower inertia implies a better cluster fit, although inertia will always fall as more centroids are added. A scree plot may be used to evaluate where inertia starts to decline at a slower pace as K increases, with the elbow representing the optimal number of centroids. A silhouette coefficient can also be used, with the highest silhouette score resulting in the optimal value of K.

LO 25.f

Natural language processing (NLP) is a component of machine learning focused on understanding and analyzing written and verbal human language. NLP offers the benefit of speed and document review without inconsistencies or bias found in human reviews.

The steps in NLP include capturing the language in a document, preprocessing the text, and analyzing it for a specific purpose. Preprocessing the text requires tokenizing the document, removing stopwords, stemming words, lemmatization, and considering N-grams.

NLP will use an inventory of sentiment words to assess whether things like corporate news releases are considered positive, negative, or neutral.

LO 25.g

Machine learning methodologies land in three categories:

- Supervised learning is used to predict either the value of a variable or the classification of an observation. The algorithm uses "labeled data" to learn.
- Unsupervised learning involves pattern recognition in the data with no specific target. This category is not used for predictions but is very useful for learning more about the data.
- Reinforcement learning incorporates a trial-and-error approach to make decisions in a changing environment.

The two primary methods used to achieve the scale consistency often required by machine learning models are standardization and normalization. Standardization is used to create a scale for measuring variables with zero mean and unit variance. Normalization is used to create a variable between zero and one which will not usually have a zero mean or unit variance.

Data cleaning is needed to manage issues such as missing data, outliers, duplicate observations, inconsistent recording, and unwanted observations.

LO 25.h

Reinforcement learning involves the creation of a policy for decision-making, with the goal of maximizing the reward. Algorithms require large quantities of training data, with an initially weak performance that will significantly improve over time.

The key areas of reinforcement learning are states, actions, and rewards. The Q-value is the expected value of taking an action (A) in a certain state (S). To determine actions taken for each state, the algorithm will choose between the best actions already identified (exploitation) and a new action (exploration).

The Monte Carlo method may be deployed to evaluate actions (A) taken in states (S) and the subsequent rewards (R) that may result. An alternative to the Monte Carlo method is the temporal difference learning method, which assumes the best strategy thus far is the one to be made going forward and will only look one decision ahead.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 25.1

1. **B** Many machine learning models require the need to scale the data used in the model. Common techniques for doing so include standardization and normalization. Relative to traditional statistical models, machine learning models provide greater flexibility, can manage large amounts of data, and can potentially capture non-linear transactions. (LO 25.a)
2. **B** Unsupervised learning is used in situations like this where a compliance manager wishes to learn more about the data but is not using it for predictive purposes. While supervised learning and reinforcement learning are established methodologies, linear regression is not an applicable machine learning category. (LO 25.g)

Module Quiz 25.2

1. **C** The goal of principal components analysis is dimensionality reduction, so a small number of components will be the output. The components should be uncorrelated, as correlated components do not independently add much value. (LO 25.d)
2. **D** The "elbow" is the point where inertia starts to decline at a slower pace as *K* increases. This represents the optimal number of centroids. A lower inertia is ideal, however, because inertia will always fall as more centroids are added, continuing to add *K* will not add value beyond a certain point. (LO 25.e)

Module Quiz 25.3

1. **C** An underfitted model excludes relevant factors and fails to capture relevant patterns. As a result, the predictions generated from such a model will have low variance but will otherwise have higher bias. (LO 25.c)
2. **D** The validation data set is used to decide between alternative machine learning models. The test set determines the effectiveness of the model once it is already chosen. The training set is used to estimate model parameters. There is no such thing as a variance set in this context. (LO 25.b)

Module Quiz 25.4

1. **B** Reinforcement learning model algorithms will choose between the best action already identified (exploitation) and new actions (exploration). Exploitation has a probability of p, which is expected to rise with additional trials. Exploration has a probability of $1 - p$ and is expected to fall with additional trials. If the exploitation probability is already at 65%, the expectation is that it will rise further as more trials are conducted. (LO 25.h)
2. **D** Stopwords are used to help sentences flow but otherwise have no value. Words like "or," "are," and "have" are considered stopwords. "Fallen" is not a stopword, as it has value in describing the direction of something (e.g., earnings, sales, etc.) (LO 25.f)

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. Cross-reference to GARP FRM Part I Quantitative Analysis, Chapter 15.

READING 26

MACHINE LEARNING AND PREDICTION

Study Session 7

EXAM FOCUS

The focus of this reading is on how machine learning models are used to generate predictions. Predictions can come from all types of models, including basic linear regression, logistic regression, and neural networks. Each model type has advantages and disadvantages relative to each other. For the exam, understand how categorical variables are encoded and the importance of model regularization using ridge regression and LASSO approaches. Methods such as decision tree construction, ensemble learning, K-nearest neighbors, and support vector machines are helpful in creating and refining models to enhance predictive value.

MODULE 26.1: CATEGORIAL VARIABLES, REGULARIZATION, AND LOGISTIC REGRESSION

Categorial Variables

LO 26.b: Understand how to encode categorical variables.

Mapping (or encoding) is the transformation of non-numerical data into numbers, which is required when using qualitative information in machine learning or regression models. Imagine a bank deciding whether to accept a loan application and the county of residence is a category, where each of the four neighboring counties is an option. Because the information does not have a natural order, a separate *dummy variable* should be set up for each category. In a process called **one-hot encoding**, an applicant would receive a 1 for their county of residence and a 0 for all other counties. For a category like income range which has a natural order (i.e., ordinal), dummy variables can take on other values. For example, in this case, a dummy variable could equal 0 for annual income less than $50,000, 1 for between $50,000 and $100,000, and 2 for above $100,000.

Regularization

LO 26.c: Discuss why regularization is useful, and distinguish between the ridge regression and LASSO approaches.

Regularization is needed for keeping models from becoming too big or complex, as it can simplify models and reduce the probability of *overfitting*. Two popular techniques for regularization are **ridge regression** and **LASSO (least absolute shrinkage and selection operator)**. Both techniques apply a penalty term to the objective (loss) function (L). By shrinking coefficient estimates in a linear regression, we can substantially reduce the variance of the model when applying test data. A third potential tool is a hybrid of ridge regression and LASSO called **elastic net**, where the loss function incorporates penalty terms from both approaches by summing them together.

Ridge Regression (L2 regularization)

Assume a dataset has the following parameters: *n* observations, *m* features, and a single output variable, *y*. Also, suppose we are estimating a linear regression model with α and β coefficients as well as a single hyperparameter, λ. The **hyperparameter** (also known as the tuning parameter) is used to select the optimal model, but it is not part of the model. An analytic approach is used to determine the values of α and β. Ridge regression looks to reduce the magnitude of the slope (β) coefficients, with the goal of shrinking them *closer to zero*.

$$L = RSS + \lambda \sum_{i=1}^{m} \beta_i^2$$

where:

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \sum_{j=1}^{m} \hat{\beta}_j x_{ij} \right)^2$$

The first component of the expression is the regression objective function (i.e., the residual sum of squares [RSS]), and the second component is the **shrinkage penalty** for large slope coefficients.

LASSO (L1 regularization)

LASSO is very similar to ridge regression, with a couple of notable differences. For ridge regression, the penalty term is the sum of squares of the slope coefficients. For LASSO, the penalty term is the sum of the absolute values of the slope coefficients. Also, a numerical (rather than analytical) approach is used to determine the α and β parameters. With this technique, slope (β) coefficients that are deemed less important are *set to exactly zero*, which means LASSO is considered a “feature selection” approach. Also, as the hyperparameter λ gets larger, a greater number of features are removed from the model.

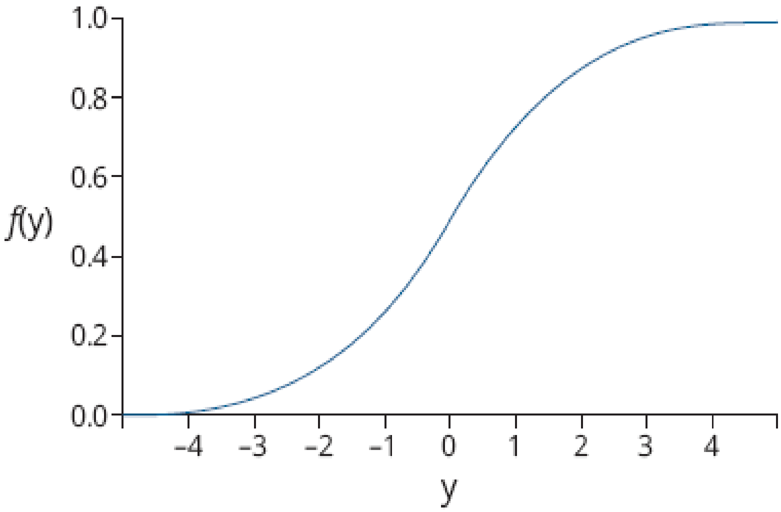
$$L = RSS + \lambda \sum_{i=1}^{m} |\beta_i|$$

Logistic Regression

LO 26.a: Explain the role of linear regression and logistic regression in prediction.

Model outputs in finance often have two potential outcomes, answering questions like whether an individual has a retirement plan, whether they have health insurance, or whether a borrower will default on a loan. In these models, there is a positive outcome (assigned a value of "1") and a negative outcome (assigned a value of "0"). While a linear model is not useful for this type of model, a **logistic regression (logit) model** can be applied. The cumulative logistic function transformation produces an output with a range from 0 to 1. The **logistic function** (also known as the sigmoid curve) is shown in Figure 26.1.

Figure 26.1: Logistic Function



The equation for this logistic function is written as:

$$f(y_j) = \frac{1}{1 + e^{-y_j}}$$

With *m* features, the functional form of y_j is written as:

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_m x_{mj}$$

The probabilities associated with y = 1 and y = 0, respectively, are:

$$y_j = 1: \quad P_j = \frac{1}{1 + e^{-(\alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_m x_{mj})}}$$

$$y_j = 0: \quad 1 - P_j$$

Because the logit model is not linear and, therefore, cannot be estimated using ordinary least squares, the **maximum likelihood method** is often used to select model parameters (α and β) to maximize the occurrence of the data training set. A **log-likelihood function** may be used instead of the likelihood function since it is easier to maximize. The log-likelihood function is written as:

$$\sum_{y_j=1} \log(f(y_j)) + \sum_{y_j=0} \log(1 - f(y_j))$$

When the parameters (α and β) have been estimated, the model can be used to create predictions by (1) setting a threshold (Z), (2) estimating the probability associated with

y = 1, and (3) specifying which category observation *j* will belong to using the following criteria:

$$\hat{y}_i = \begin{cases} 1 & \text{if } P_j \geq Z \\ 0 & \text{if } P_j < Z \end{cases}$$

Setting a value for *Z* will depend on the costs of being wrong on both sides. A *Z* equal to 0.5 implies that the costs of being wrong (i.e., classifying a value of *y* as zero when it should be one, or vice versa) are equal. For example, a homeowner defaulting (y = 1) on a mortgage when the bank assumes they will pay back is far costlier than a homeowner paying (y = 0) the mortgage when the bank assumes they will default (and therefore doesn't make the loan); as such, *Z* may be set to a low number such as 0.1.

Regarding model evaluation, if the model's output is a continuous variable such as a rate of return, the **mean squared forecast error (MSFE)** can be calculated for the data testing set as follows:

$$\text{MSFE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2$$

The **mean absolute forecast error** is an alternative approach that uses absolute values instead of squares.

MODULE QUIZ 26.1

1. A financing company uses household income ranges to model loan applications. The ranges are in \$25,000 increments up to \$200,000, and then the last range is anything above \$200,000. If a family has a household income of \$85,000, the dummy variable most likely assigned will be:
   A. 0.
   B. 1.
   C. 2.
   D. 3.

2. Which of the following statements about shrinkage penalty terms in regularization is most accurate?
   A. Penalty terms are not applied in model regularization.
   B. The sum of the squares of the slope coefficients is the penalty term for LASSO.
   C. Elastic net sums the penalty terms found in both ridge regression and LASSO.
   D. The sum of the absolute values of the slope coefficients is the penalty term for ridge regression.

3. For a given logistic regression model, default is assigned a value of 1 and no default is assigned a value of 0. Because the cost of default is very high on loans the bank expects its consumers to pay back, a reasonable threshold level for *Z* should be:
   A. 0.00.
   B. 0.10.
   C. 0.50.
   D. 0.90.

MODULE 26.2: DECISION TREES, ENSEMBLE LEARNING, K-NEAREST NEIGHBORS, AND SUPPORT VECTOR MACHINES

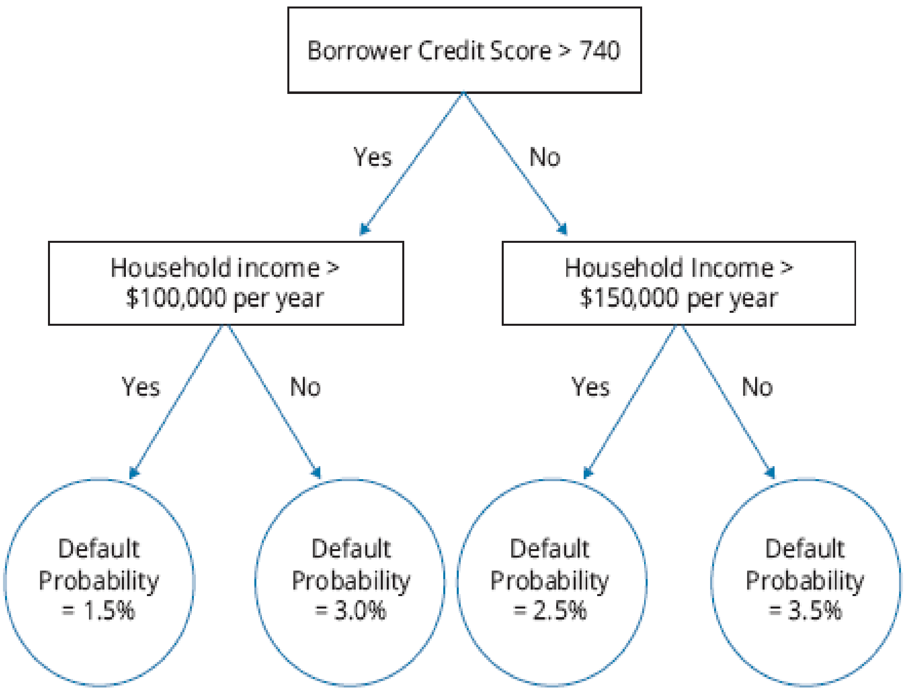
Decision Trees

LO 26.d: Show how a decision tree is constructed and interpreted.

Decision trees are supervised machine learning techniques that visually represent a tree and work with sequential input features. Every node of the tree has a question that reflects an observation that is connected to another node (leaf) by a branch. Decision trees are useful for classification problems and continuous variable estimations and, as such, are sometimes referred to as **classification and regression trees (CARTs)**. Because CARTs are easy to interpret, they are sometimes referred to as "white-box models" as opposed to neural networks, which are considered "black-box models."

Figure 26.2 shows an example of a decision tree that a bank may use to evaluate the default probability of a potential borrower. The top node in the tree is known as the *root node*. Internal nodes are known as *decision nodes*, and leaf nodes are known as *terminal nodes*.

Figure 26.2: Decision Tree Example



Information gain measures the extent to which obtaining information about a given feature can reduce uncertainty. In each node, we are looking for the feature that maximizes information gain. Entropy and the Gini coefficient are popular measures of information gain.

- **Entropy** is a measure of disorder in a dataset whose value lies between 0 and 1. The equation for entropy, where *n* is the number of possible outcomes and *p* is the probability of each outcome, is written as: entropy = $-\sum_{i=1}^{n} p_i \log_2(p_i)$
- The **Gini measure** (also known as Gini impurity) is written as:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

To illustrate how information gain is measured, assume a financial analyst builds a model to predict whether a company will meet its interest obligations during the next

fiscal year. Meeting all interest obligations equates to a value of y = 1, and not meeting all obligations (whether in terms of timing or amount) equates to a value of y = 0.

As shown in Figure 26.3, the variables included in the model are whether sales have increased or decreased in the previous year, the percentage of previous year sales made on credit versus cash, whether the bond rating is investment grade or below investment grade, and whether earnings have increased or decreased in the previous year. An output value of 1 is assigned if sales have increased, the bond rating is investment grade, and earnings have increased. An output value of 0 is assigned if sales have decreased, the bond rating is below investment grade, and earnings have decreased.

Figure 26.3: Decision Tree Data

Data Point	Interest	Sales Change	Credit Sales	Bond Rating	Earnings Change
1	1	1	60	1	1
2	0	0	70	1	0
3	0	1	50	1	1
4	1	1	70	0	1
5	1	0	80	0	1
6	0	0	60	1	0
7	1	0	70	0	1
8	0	1	60	1	1
9	1	1	70	1	0
10	1	0	70	0	1
11	0	1	50	1	0
12	1	1	80	1	1
13	0	1	70	0	0
14	1	0	80	0	1
15	1	1	90	1	0
16	1	0	70	0	1

In an ideal situation, a decision tree question in root and decision nodes produces a perfect split (a "pure set"). For example, a pure set occurs when all companies with earnings reductions did not meet interest obligations. The opposite is a situation where half of the companies with earnings reductions met interest obligations and the other half did not, which would make the earnings information essentially meaningless.

Based on the output shown in Figure 26.3, 10 of the 16 firms met their *interest obligations* and 6 did not. The Gini coefficient for this output is computed as follows:

$$G = 1 - \left\{ \left(\frac{10}{16}\right)^2 + \left(\frac{6}{16}\right)^2 \right\} = 0.46875$$

This number serves as the base level to compare the drop in the Gini coefficient as the tree gets larger. The selected root node of the tree will be the variable that causes the greatest decline in the Gini coefficient (i.e., the highest information gain). As an example, the calculation of the information gained from the *sales change* is shown as follows:

9 firms had sales increases, and 5 of those firms met interest obligations and 4 did not:

$$G = 1 - \left\{ \left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2 \right\} = 0.49383$$

7 firms had sales decreases, and 5 of those firms met interest obligations and 2 did not:

$$G = 1 - \left\{ \left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2 \right\} = 0.40816$$

$$\text{weighted Gini} = \frac{9}{16} \times 0.49383 + \frac{7}{16} \times 0.40816 = 0.45635$$

Therefore, the information gain is equal to the base Gini measure of 0.46875 minus the weighted Gini measure of 0.45635. This results in an information gain of 0.0124.

For information like the *percentage of sales on credit*, because it is a continuous variable that can take any value from zero to one hundred, the threshold value chosen will be the one which maximizes the information gain.

The decision tree is complete when all features have been used or when a leaf is reached that is a pure set. A key risk of decision trees is overfitting, which can be mitigated by setting stopping rules such that there is a maximum number of branches. *Pre-pruning* occurs when splitting stops if the training set observation count relating to a node is under a specific number. *Post-pruning* occurs when a large tree is built, and weak nodes are removed.

Ensemble Learning

LO 26.e: Describe how ensembles of learners are built.

To construct ensembles of learners, the outputs of a range of different models are combined into a single model. The two objectives of this outcome are the benefits of averaging many predictions (i.e., the wisdom of crowds) and protection against overfitting. Decision trees are one of the many machine learning model types that can be used to build an ensemble of learners. Three techniques for building ensembles are described as follows:

- **Bootstrap aggregation** (or **bagging**): multiple decision trees are created from among the training sample and the classifications or predictions from the trees are aggregated to create a new classification or prediction. Taking the average of all bagged trees reduces the variance of the model. Data is sampled with replacement, which means some observations may never appear. The "out-of-bag" observations can then be used to evaluate the performance of the model. *Pasting* differs from bagging in that the former involves sampling with no replacement, such that 500 items in a training set with sub-samples of 25 items would equate to 20 sub-samples.
- **Random forests**: an ensemble of decision trees which is created by sampling features or observations without replacement. This process is repeated hundreds

of times for a “forest” of decision trees. Random forests apply bagging techniques but improve upon them by reducing the correlations between decision trees. The number of features used is often equivalent to the square root of total model predictors available. Model outputs with low correlations produce the greatest performance for ensembles.

- **Boosting**: designed to improve model performance based on previously grown trees. Boosting consists of gradient boosting and adaptive (*AdaBoost*) boosting. The former builds a new model using the residuals of the previous model, while the latter involves a model with equal weights on all observations with sequential weight increases on misclassified outputs.

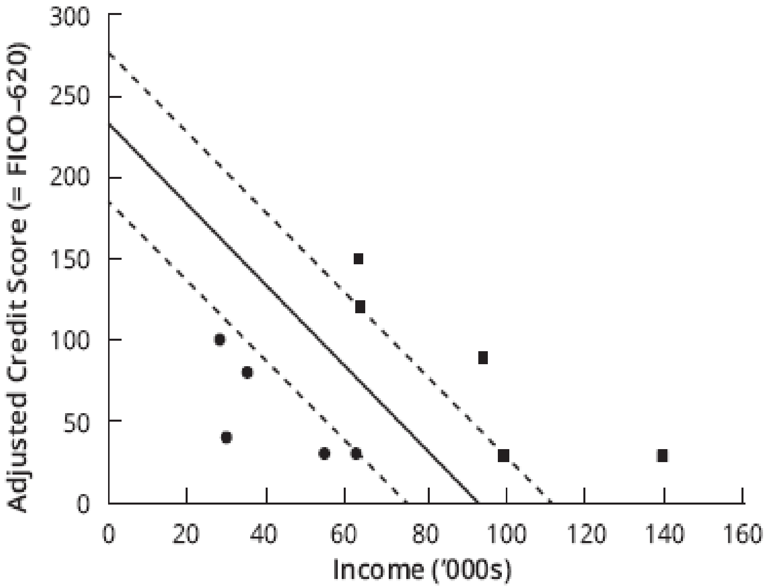
K-Nearest Neighbors and Support Vector Machines

LO 26.f: Outline the intuition behind the K nearest neighbors and support vector machine methods for classification.

K-nearest neighbors (KNN) is supervised machine learning model used to classify or predict the value of a target variable. While other approaches learn dataset relationships, KNN does not and is therefore considered a “lazy learner.” The KNN implementation involves selecting a K value and distance measure and for each training sample data point, identifying the feature space K closest observations to the point which is being predicted. When predicting a target value, the target can be set equal to the average of its values for the nearest neighbors. The value of K is critical, as a K that is too large (small) will produce a high (low) bias and low (high) variance. The value for K is often set equal to the square root of the training sample size (n), such that n of 400 points results in a K of 20.

Support vector machines (SVMs) are supervised machine learning models which are beneficial when there is a large quantity of features. In a simplified example where assessing borrower default incorporates only two features (e.g., credit score and income), the goal of SVMs is to use these features to develop a line that graphically separates the two groups (e.g., default versus no default). SVMs create the widest path using two parallel lines to separate the different observation classes. Support vectors are the data points lying on the edge of the paths, while the separation boundary represents the center of the path.

Figure 26.4: Support Vector Machine Example



Although a model with two features is the most basic, the optimization framework and underlying principles are the same regardless of how many features are modeled. The output would be a *hyperplane* with the dimension count equal to the number of features minus one. Beyond a two-feature model, there will be a tradeoff between the path width and the extent of path-driven misclassifications.

MODULE QUIZ 26.2

1. The Gini coefficient for a model is 0.375. If the weighted Gini of one of the features is 0.329, the information gain will be closest to:
   A. 0.046.
   B. 0.352.
   C. 0.704.
   D. 0.750.
2. In relation to ensembles of learners, which of the following statements best describes the "wisdom of crowds"?
   A. Masses of people are never wrong.
   B. It is only evident using decision trees.
   C. It protects against over or underfitting.
   D. It offers the benefits of averaging many predictions.
3. In a two-feature support vector machine, the separation boundary is best described as the:
   A. center of the path.
   B. the lower bound of the path.
   C. the upper bound of the path.
   D. all data points lying on both edges of the path.

MODULE 26.3: NEURAL NETWORKS AND PERFORMANCE

Neural Networks

LO 26.g: Understand how neural networks are constructed and how their weights are determined.

Similar to how the human brain works to perform calculations, **artificial neural networks (ANNs)** are machine learning approaches used for computations. The feedforward network with backpropagation is the most common type of ANN. Backpropagation is a description of how biases and weights are constantly updated through model iterations.

Figure 26.5: Neural Network Example

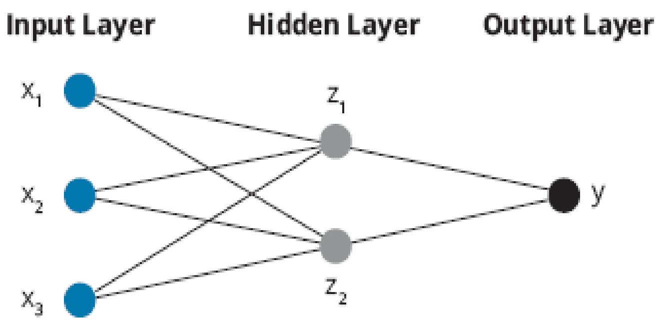


Figure 26.5 illustrates a feedforward network with three features (input variables), a single output variable (*y*), and one hidden layer with two nodes. To determine the value of hidden layer nodes, weights (*w*) are applied to the inputs. A constant **bias** (i.e., activation parameter) is then added, which represents how easy it is to get a node to “fire” (i.e., generate an output of 1). Weights and biases are combined to generate output for each node as follows:

$$\text{output} = \text{bias} + \sum_{j=1}^{m} w_j x_j$$

Each layer will also have its own **activation function**. A common activation function for neural networks is the logistic function discussed earlier. The goal of a neural network is to identify non-linear relationships, and activation functions are used to introduce non-linearity into input and output relationships.

Because there is no analytical formula for selecting the best parameter values, the **gradient descent algorithm** may be applied to minimize the objective (loss) function. Trial values of the parameters are chosen and then the direction the values should change to improve the objective function value. This is referred to as the “line of steepest descent down a valley” of the function. The **learning rate** is a hyperparameter that reflects the size of the steps taken. While a learning rate that is too large will produce large movements between sides of the valley, a learning rate which is too small will produce a gradient descent algorithm that takes too long to find the global minimum value.

Overfitting is a concern when a neural network has many hidden layers and nodes per layer. Performing calculations for the validation and training datasets at the same time can eliminate overfitting. Moving down the valley will improve both dataset objective functions, but the place to stop the gradient descent algorithm is the point where the objective function value declines for the validation set even as it continues to improve for the training set.

Predictive Performance

LO 26.h: Evaluate the predictive performance of logistic regression models and neural network models using a confusion matrix.

When a model has a binary categorical output (i.e., 0 or 1), a **confusion matrix** may be used to evaluate the model. A confusion matrix commonly has a 2 × 2 format and shows possible outcomes along with reflecting the accuracy of prediction. The four elements of the matrix are:

1. **True positives (TP)**: a positive outcome was predicted by the model, and it was actually true.
2. **False positives (FP)**: a positive outcome was predicted by the model, but it was actually false.
3. **True negatives (TN)**: a negative outcome was predicted by the model, and it was actually false.
4. **False negatives (FN)**: a negative outcome was predicted by the model, but it was actually true.

For illustration purposes, assume a model is built to determine the probability that a company will acquire another company. The model is based on 500 companies, where 150 companies made acquisitions and 350 did not.

Figure 26.6: Confusion Matrix

		Model Prediction	
		Acquisition	No acquisition
Outcome/Actual	Acquisition	96 (TP) = 19.2%	54 (FN) = 10.8%
	No acquisition	124 (FP) = 24.8%	226 (TN) = 45.2%

The following performance metrics may be derived from the confusion matrix:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{96 + 226}{96 + 226 + 124 + 54} = 64.4\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{96}{96 + 124} = 43.6\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{96}{96 + 54} = 64.0\%$$

$$\text{Error Rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{124 + 54}{96 + 226 + 124 + 54} = 35.6\%$$

A **receiver operating characteristic (ROC) curve** is a means of illustrating the link between true positive rates and false positive rates. The model predictions improve as the **area under the ROC curve (AUC)** increases.

- AUC = 1: completely accurate predictions
- AUC = 0.5: model has no predictive value (i.e., random model)
- AUC < 0.5: model has negative predictive value

Figure 26.7: ROC Curve Example

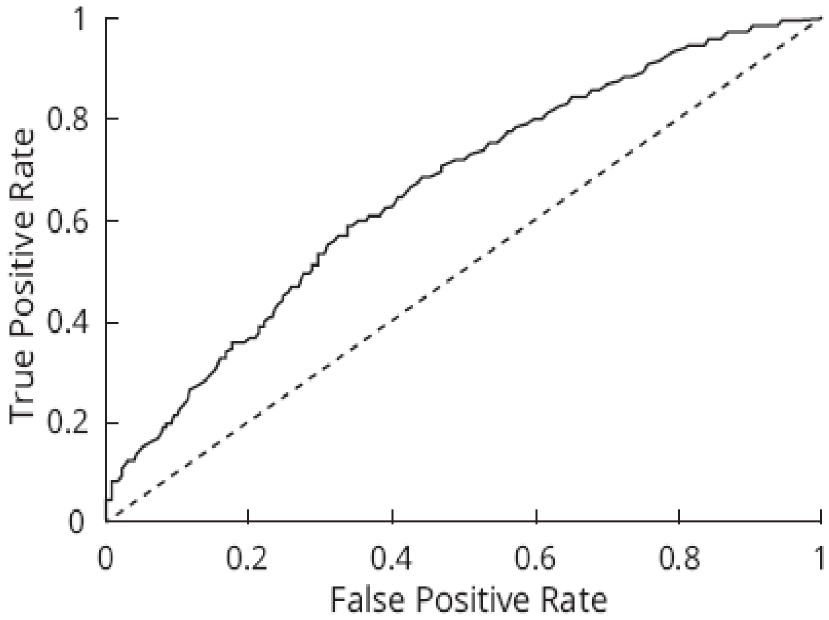


Figure 26.8 shows a hypothetical comparison of logistic regression and neural network models for the assessment of customer defaults for bank loans.

Figure 26.8: Logistic Regression and Neural Network Model Comparisons

	Training Sample		Validation Sample	
Measure	Logistic Regression	Neural Network	Logistic Regression	Neural Network
Accuracy	0.775	0.775	0.645	0.620
Precision	0.612	0.659	0.582	0.557
Recall	0.297	0.228	0.345	0.380

An assessment of the data shows that the model fit is stronger on the training data than it is on the validation data, which reflects slight overfitting. In looking at the validation sample, the accuracy and precision were superior for the logistic regression, but recall was superior for the neural network. For the training sample, accuracy was equal across both methods, but precision was superior for the neural network and recall was superior for the logistic regression. These indicators are considered contradictory because no obvious conclusion as to which model is better can be derived.

A confusion matrix can be used to assess the raw data and may better reflect the divergence summarized in Figure 26.8. One model may predict more defaults for one sample, while the other model may predict more defaults for the other sample. The set with the highest true positive and true negative rates would therefore differ between the models. Figure 26.9 reflects hypothetical confusion matrices for logistic regression and neural network validation and training samples.

Figure 26.9: Confusion Matrix for Model Comparisons

Logistic Regression Training Sample			
		Prediction	
		No default	Default
Outcome	No default	322	10
	Default	56	18

Logistic Regression Validation Sample			
		Prediction	
		No default	Default
Outcome	No default	95	8
	Default	30	14

Neural Network Training Sample			
		Prediction	
		No default	Default
Outcome	No default	82	13
	Default	20	16

Neural Network Validation Sample			
		Prediction	
		No default	Default
Outcome	No default	374	7
	Default	70	12

MODULE QUIZ 26.3

1. Which of the following statements best describes when to stop the gradient descent algorithm in a neural network model?
   A. When the value for the objective function increases for both the validation set and the training set.
   B. When the value for the objective function decreases for both the validation set and the training set.
   C. When the value for the objective function declines for the training set, even as it improves for the validation set.
   D. When the value for the objective function declines for the validation set, even as it improves for the training set.

2. In a confusion matrix established for a logistic regression model, which two performance metrics must sum to 100%?
   A. Precision and recall.
   B. Recall and error rate.
   C. Accuracy and error rate.
   D. Accuracy and precision.

3. A confusion matrix on a logistic regression model shows 35 true positives, 28 false negatives, 12 false positives, and 25 true negatives. The precision metric will show a percentage output of:
   A. 40%.
   B. 56%.
   C. 60%.
   D. 74%.

KEY CONCEPTS

LO 26.a

Model outputs in finance often have two potential outcomes. In these models, there is a positive outcome (assigned a value of "1") and a negative outcome (assigned a value of "0"). Because a linear regression model will not work for this type of modeling, a logistic regression (logit) model can be used instead which produces an output with a range of 0 to 1.

When the parameters (α and β) have been estimated, the model can be used to create predictions by (1) setting a threshold (Z), (2) estimating the probability associated with y = 1, and (3) specifying the category in which the observation *j* will land. Setting a value for *Z* will depend on the costs of being wrong on both sides. A *Z* equal to 0.5 implies that the costs of being wrong (i.e., classifying a value of *y* as zero when it should be one, or vice versa) are equal. The *Z* will shift depending on where there is cost imbalance associated with each outcome.

LO 26.b

Mapping (or encoding) is the transformation of non-numerical data into numbers. For information that does not have a natural order, a separate dummy variable (0 or 1) should be set up for each category. For a category that has a natural order, dummy variables can extend beyond 0 and 1.

LO 26.c

Regularization is needed to simplify models and reduce the probability of overfitting. Two techniques for regularization are ridge regression and LASSO (least absolute shrinkage and selection operator), which both apply a penalty term to the objective (loss) function. A third potential tool is a hybrid of ridge regression and LASSO called elastic net, where the loss function incorporates both penalty terms by summing them together. For both approaches, the first component of the expression is the regression objective function (i.e., residual sum of squares [RSS]) and the second component is the shrinkage penalty term for large slope parameters.

For ridge regression, the penalty term is the sum of the squares of the slope coefficients. For LASSO, the penalty term is the sum of the absolute values of the slope coefficients. Ridge regression uses an analytical approach for its α and β parameters, while LASSO uses a numerical approach. For ridge regression, the goal is to shrink slope coefficients close to zero. For LASSO, the slope coefficients that are deemed less important are set to exactly zero.

LO 26.d

Decision trees are supervised machine learning techniques that visually represent an upside-down tree and work with sequential input features. Every node of the tree has a question that reflects an observation connected to another node (leaf) by a branch.

Information gain measures the extent to which obtaining information about a given feature can reduce uncertainty. In each node, we are looking for the feature that maximizes information gain. Entropy and the Gini coefficient are used as ways to compute information gain.

The decision tree is complete when all features have been used or when a leaf is reached that is a pure set. Pre-pruning is when splitting stops if the training set observation count relating to a node is under a specific number. Post-pruning is when a large tree is built, and weak nodes are removed.

LO 26.e

Ensemble learning provides the benefit of averaging many predictions and protects against overfitting. Three techniques for building ensembles are bootstrap aggregation (bagging), random forests, and boosting.

LO 26.f

K-nearest neighbors (KNN) is a supervised machine learning model used to classify or predict the value of a target variable. The value of K is critical, as a value that is too large (small) will produce a high (low) bias and low (high) variance. K is often set as the square root of the training sample size.

Support vector machines (SVMs) are supervised machine learning models which are beneficial when there is a large quantity of features. In a simplified two-feature model, SVM creates the widest path using two parallel lines which separate the observations. Support vectors are the data points lying on the edge of the path, while the separation boundary represents the center of the path. The optimization framework and underlying principles are the same regardless of how many features are modeled, although the output would be a hyperplane with the dimension count equal to the number of features minus one.

LO 26.g

Artificial neural networks (ANNs) are machine learning approaches used for computations. Feedforward networks with backpropagation are the most common type of ANN. Backpropagation is a description of how biases and weights are constantly updated through model iterations.

The goal of a neural network is to identify non-linear relationships. The gradient descent algorithm may be applied to minimize the objective (loss) function. The learning rate is a hyperparameter that reflects the size of the steps taken as the model goes down the line of steepest descent down a valley of the function. While a learning rate that is too large will produce large movements between sides of the valley, a learning rate which is too small will produce a gradient descent algorithm that takes too long to find the global minimum value.

Overfitting is a concern when a neural network has many hidden layers and nodes per layer. Performing calculations for the validation and training datasets at the same time can eliminate overfitting. The place to stop the gradient descent algorithm is the point where the objective function value declines for the validation set even as it continues to improve for the training set.

LO 26.h

When a model has a binary categorical output (i.e., 0 or 1), a confusion matrix may be used to evaluate the model. A confusion matrix is useful for logistic regression models and neural network models. The four elements of the matrix are:

1. True positives (TP): a positive outcome was predicted by the model, and it was actually true.
2. False positives (FP): a positive outcome was predicted by the model, but it was actually false.
3. True negatives (TN): a negative outcome was predicted by the model, and it was actually false.
4. False negatives (FN): a negative outcome was predicted by the model, but it was actually true.

Performance metrics derived from the confusion matrix include accuracy, precision, recall, and error rate. A receiver operating characteristic (ROC) curve is a means of

illustrating the link between true positives and false positives. The model predictions are improved as the area under the ROC curve (AUC) increases.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 26.1

1. **D** This category has a natural order, so the dummy variables will be assigned (starting with 0) for each consecutive range. Dummy variable 0 is for income from $0–$25,000, 1 for income from $25,000–$50,000, 2 for income from $50,000–$75,000, and 3 for income from $75,000–$100,000. $85,000 in household income falls within the latter range, which means the dummy variable will be set to 3. (LO 26.b)

2. **C** Elastic net is a regularization technique where the loss function incorporates penalty terms from ridge regression and LASSO. The sum of the squares of the coefficients is the penalty term for ridge regression, while the sum of the absolute values of the coefficients is the penalty term for LASSO. (LO 26.c)

3. **B** The *Z* value should be low to account for the asymmetry in risk associated with a loan that defaults (when unexpected) versus when loan payments are met (when unexpected). Setting a value equal to zero is unrealistic, and a value equal to 0.50 implies risk symmetry (which is not the case here). So, a *Z* of 0.10 is the most appropriate choice in this situation. (LO 26.a)

Module Quiz 26.2

1. **A** The information gain is the difference between the base Gini and the weighted Gini. For this model, the base Gini is 0.375 and the weighted Gini is 0.329. The difference is equal to 0.375 – 0.329 = 0.046. (LO 26.d)

2. **D** The wisdom of crowds is evident through ensembles of learners, where many different model predictions are made, and they can be averaged to derive a best estimate proxy. While individual models on their own are vulnerable to error, predictions from multiple models can be averaged to produce the best estimate. (LO 26.e)

3. **A** In a two-feature support vector machine, the widest path using two parallel lines which separate the observations is created. Support vectors are the data points lying on the edge of the path, while the separation boundary represents the center of the path. (LO 26.f)

Module Quiz 26.3

1. **D** To prevent overfitting, performing calculations for the validation and training datasets at the same time can be done. Moving down the valley will improve both dataset objective functions, but the point to stop the gradient descent algorithm is the point where the objective function value declines for the validation set even as it continues to improve for the training set. (LO 26.g)

2. **C** The numerator of the accuracy metric captures true positives and true negatives. The numerator of the error rate captures false positives and false negatives. Both metrics have all four outcomes in the denominator, which implies that when the metrics are added together, all four outcomes are in both the numerator and denominator. The sum must therefore be 100%. (LO 26.h)

3. **D** Precision is equal to true positives (35) divided by the sum of true positives and false positives (35 + 12, or 47). 35/47 = 74%. There are 100 total outcomes (35 + 28 + 12 + 25). The error rate is the “false” outcomes (28 + 12) divided by total outcomes or 40%. Recall is the true positives (35) divided by the sum of the true positives and the false negatives (35 + 28, or 63). 35/63 = 56%. Accuracy is the “true” outcomes (35 + 25) divided by total outcomes or 60%. (LO 26.h)

FORMULAS

Reading 12

joint probability: $P(AB) = P(A \mid B) \times P(B)$

conditional probability: $P(A \mid B) = \frac{P(AB)}{P(B)}$

Bayes' formula: $P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$

Reading 13

expected value: $E(X) = \Sigma P(x_i)x_i = P(x_1)x_1 + P(x_2)x_2 + \ldots + P(x_n)x_n$

variance: $\sigma^2 = E\{[X - E(X)]^2\} = E[(X - \mu)^2]$

$$\text{skewness} = \frac{E\left[(X - \mu)^3\right]}{\sigma^3}$$

$$\text{kurtosis} = \frac{E\left[(X - \mu)^4\right]}{\sigma^4}$$

Reading 14

uniform distribution range: $P(x_1 \le X \le x_2) = (x_2 - x_1) / (b - a)$

PDF of continuous uniform distribution: $f(x) = \frac{1}{b - a}$ for $a \le x \le b$, else $f(x) = 0$

mean of uniform distribution: $E(x) = \frac{a + b}{2}$

variance of uniform distribution: $\text{Var}(x) = \frac{(b - a)^2}{12}$

binomial probability function: $p(x) = \frac{n!}{(n - x)!x!} p^x (1 - p)^{n-x}$

expected value of binomial random variable: expected value of $X = E(X) = np$

variance of binomial random variable: variance of $X = np(1 - p)$

Poisson distribution: $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

chi-squared test statistic:

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2}$$

where:
n = sample size
s^2 = sample variance
σ_0^2 = hypothesized value for the population variance

F-test:

$$F = \frac{s_1^2}{s_2^2}$$

where:
s_1^2 = variance of the sample of n_1 observations drawn from Population 1
s_2^2 = variance of the sample of n_2 observations drawn from Population 2

Reading 15

covariance:

$$Cov[X_1,X_2] = E[(X_1 - E[X_1])(X_2 - E[X_2])]$$

$$Cov[X_1,X_2] = E[X_1X_2] - E[X_1]E[X_2]$$

correlation: $Corr(X_1,X_2) = \dfrac{Cov[X_1,X_2]}{\sqrt{Var[X_1]}\sqrt{Var[X_2]}} = \dfrac{\sigma_{12}}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}} = \dfrac{\sigma_{12}}{\sigma_1\sigma_2}$

variance of a two-asset portfolio: $\sigma_{12}^2 = w_1^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w_1(1-w)\sigma_{12}$

Reading 16

sample mean: $\hat{\mu} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

sample variance: $s^2 = \dfrac{n}{n-1}\hat{\sigma}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2$

sample covariance: sample $Cov_{XY} = \dfrac{\sum_{i=1}^{n}\left(X_i - \hat{\mu}_X\right)\left(Y_i - \hat{\mu}_Y\right)}{n-1}$

Reading 17

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the sample statistic}}$$

$$\text{standard error of sample mean: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

confidence interval:

$$\left\{\left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} - \left(\begin{matrix}\text{critical}\\\text{value}\end{matrix}\right)\left(\begin{matrix}\text{standard}\\\text{error}\end{matrix}\right)\right] \leq \begin{matrix}\text{population}\\\text{parameter}\end{matrix} \leq \left[\begin{matrix}\text{sample}\\\text{statistic}\end{matrix} + \left(\begin{matrix}\text{critical}\\\text{value}\end{matrix}\right)\left(\begin{matrix}\text{standard}\\\text{error}\end{matrix}\right)\right]\right\}$$

$$z\text{-statistic} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

where:
$\bar{x}$ = sample mean
μ_0 = hypothesized population mean
σ = standard deviation of the population
n = sample size

$$\text{equality of means test statistic: } T = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{s_X^2 + s_Y^2 - 2\text{cov}(X, Y)}{n}}}$$

Reading 18

regression equation:

$$Y = \alpha + \beta \times (X) + \varepsilon$$

where:
β = the regression or slope coefficient; sensitivity of Y to changes in X
α = value of Y when X = 0
ε = random error or shock; unexplained (by X) component of Y

$$\text{regression slope coefficient: } \beta = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

regression intercept:

$$\alpha = \bar{Y} - \beta\bar{X}$$

where:
$\bar{Y}$ = mean of Y
$\bar{X}$ = mean of X

confidence interval of the slope coefficient = $\beta \pm (t_c \times S_b)$

Reading 19

multiple regression:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where:

β_j = the regression or slope coefficients; sensitivity of Y to changes in X_j controlling for all other Xs

α = value of Y when all Xs = 0

ε = random error or shock; unexplained (by X) component of Y (This error may be reduced by using more independent variables or by using different, more appropriate independent variables.)

total sum of squares:

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum\left(Y_i - \hat{Y}\right)^2$$

or

TSS = ESS + RSS

where:

TSS = total sum of squares (i.e., total variation in Y)

ESS = explained sum of squares (i.e., the variation in Y explained by the regression model)

RSS = residual sum of squares (i.e., the unexplained variation in Y)

coefficient of determination: R^2 = ESS/TSS

adjusted R^2:

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times \left(1 - R^2\right)\right]$$

where:

n = number of observations

k = number of independent variables

F-statistic:

$$F = \frac{\left(RSS_P - RSS_F\right)/q}{RSS_F/\left(n - k_F - 1\right)} = \frac{\left(R_F^2 - R_P^2\right)/q}{\left(1 - R_F^2\right)/\left(n - k_F - 1\right)}$$

where:

RSS_F = residual sum of squares of the full model

RSS_P = residual sum of squares of the partial model

R_F^2 = coefficient of determination of the full model

R_P^2 = coefficient of determination of the partial model

q = number of restrictions imposed on the full model to arrive at the partial model

n = number of observations

k_F = number of independent variables in the full model

Reading 20

variance inflation factor: $VIF_j = \frac{1}{1 - R_j^2}$

Cook's measure:

$$D_j = \frac{\sum_{i=1}^{n} (\hat{y}_i^{(-j)} - \hat{y}_i)^2}{kS^2}$$

where:
$\hat{y}^{(-j)}$ = predicted value of *y* after dropping outlier observation *j*
$\hat{y}_i$ = predicted value of *y* without dropping any observation
k = number of independent variables
S^2 = squared residuals in the model with all observations

Reading 21

first-order autoregressive [AR(1)] process:

$$y_t = d + \Phi y_{t-1} + \varepsilon_t$$

where:
d = intercept term
y_t = the time series variable being estimated
y_{t-1} = one-period lagged observation of the variable being estimated
ε_t = current random white noise shock (mean 0)
Φ = coefficient for the lagged observation of the variable being estimated

first-order moving average [MA(1)] process:

$$y_t = \mu + \theta\varepsilon_{t-1} + \varepsilon_t$$

where:
μ = mean of the time series
ε_t = current random white noise shock (mean 0)
ε_{t-1} = one-period lagged random white noise shock
θ = coefficient for the lagged random shock

lag operator: $y_{t-1} = Ly_t$

autoregressive moving average (ARMA) process:

$$y_t = d + \Phi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$$

where:
d = intercept term
y_t = the time series variable being estimated
Φ = coefficient for the lagged observations of the variable being estimated
y_{t-1} = one-period lagged observation of the variable being estimated
ε_t = current random white noise shock
θ = coefficient for the lagged random shocks
ε_{t-1} = one-period lagged random white noise shock

Reading 22

linear time trend: $y_t = \delta_0 + \delta_1 t + \varepsilon_t$

nonlinear time trend: $y_t = \delta_0 + \delta_1 + \delta_2 t^2 + \varepsilon_t$

log-quadratic model: $\ln(y_t) = \delta_0 + \delta_1 t + \delta_2 t^2 + \varepsilon_t$

pure seasonal dummy variable model: $y_t = \sum_{i=1}^{s} \gamma_i \left(D_{i,t}\right) + \varepsilon_t$

trend model with seasonality: $y_t = \beta_1(t) + \sum_{i=1}^{s} \gamma_i \left(D_{i,t}\right) + \varepsilon_t$

Reading 23

asset return: $R_t = \dfrac{P_t - P_{t-1}}{P_{t-1}}$

continuously compounded asset return: $r_t = \ln P_t - \ln P_{t-1}$

annualized volatility: $\sigma_{annual} = \sqrt{252 \times \sigma^2_{daily}}$

Jarque-Bera test statistic: $JB = (T - 1)\left(\dfrac{\hat{S}^2}{6} + \dfrac{\left(\hat{K} - 3\right)^2}{24}\right)$

power law: $P(X > x) = kx^{-\alpha}$

Spearman's rank correlation: $\widehat{\rho_s} = 1 - \dfrac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2 - 1)}$

Kendall's τ (tau):

$$\hat{\tau} = \frac{n_c - n_d}{n(n-1)/2} = \frac{n_c}{n_c + n_d + n_t} - \frac{n_d}{n_c + n_d + n_t}$$

where:
n_c = number of concordant pairs
n_d = number of discordant pairs
n_t = number of ties

Reading 25

Standardization:

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \hat{\mu}_i}{\hat{\sigma}_i}$$

where:

$\hat{\mu}_i$ = estimated mean

$\hat{\sigma}_i$ = estimated standard deviation

Normalization:

$$\tilde{x}_{i,j} = \frac{x_{i,j} - x_{i,min}}{x_{i,max} - x_{i,min}}$$

where:

$x_{i,min}$ = minimum of the observations

$x_{i,max}$ = maximum of the observations

Euclidean distance:

Two features: $d_E = \sqrt{(x_{1Q} - x_{1P})^2 + (x_{2Q} - x_{2P})^2}$

m features: $d_E = \sqrt{\sum_{i=1}^{m}(x_{iQ} - x_{iP})^2}$

Manhattan distance:

Two features: $d_M = |x_{1Q} - x_{1P}| + |x_{2Q} - x_{2P}|$

m features: $d_M = \sum_{i=1}^{m}|x_{iQ} - x_{iP}|$

$$\text{inertia} = \sum_{j=1}^{n} d_j^2$$

Reading 26

Ridge regression:

$$L = RSS + \lambda \sum_{i=1}^{m} \beta_i^2$$

where:

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \sum_{j=1}^{m} \hat{\beta}_j x_{ij} \right)^2$$

LASSO regression:

$$L = RSS + \lambda \sum_{i=1}^{m} |\beta_i|$$

Logistic regression function:

$$f(y_j) = \frac{1}{1 + e^{-y_j}}$$

Entropy:

$$\text{entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Gini coefficient:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

APPENDIX

USING THE CUMULATIVE *Z*-TABLE

Probability Example

Assume that the annual earnings per share (EPS) for a large sample of firms is normally distributed with a mean of \$5.00 and a standard deviation of \$1.50. What is the approximate probability of an observed EPS value falling between \$3.00 and \$7.25?

If EPS = x = \$7.25, then z = (x − μ) / σ = (\$7.25 − \$5.00) / \$1.50 = +1.50

If EPS = x = \$3.00, then z = (x − μ) / σ = (\$3.00 − \$5.00) / \$1.50 = −1.33

For z*-value of 1.50:* Use the row headed 1.5 and the column headed 0 to find the value 0.9332. This represents the area under the curve to the left of the critical value 1.50.

For z*-value of −1.33:* Use the row headed 1.3 and the column headed 3 to find the value 0.9082. This represents the area under the curve to the left of the critical value +1.33. The area to the left of −1.33 is 1 − 0.9082 = 0.0918.

The area between these critical values is 0.9332 − 0.0918 = 0.8414, or 84.14%.

Hypothesis Testing—One-Tailed Test Example

A sample of a stock's returns on 36 non-consecutive days results in a mean return of 2.0%. Assume the population standard deviation is 20.0%. Can we say with 95% confidence that the mean return is greater than 0%?

H_0: μ ≤ 0.0%, H_A: μ > 0.0%. The test statistic = *z*-statistic = $\dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
= (2.0 − 0.0) / (20.0 / 6) = 0.60.

The significance level = 1.0 − 0.95 = 0.05, or 5%.

Since this is a one-tailed test with an alpha of 0.05, we need to find the value 0.95 in the cumulative *z*-table. The closest value is 0.9505, with a corresponding critical *z*-value of 1.65. Since the test statistic is less than the critical value, we fail to reject H_0.

Hypothesis Testing—Two-Tailed Test Example

Using the previous assumptions, suppose that the analyst now wants to determine with 99% confidence that the stock's return is not equal to 0.0%.

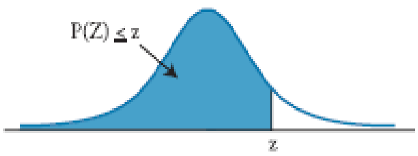
H_0: μ = 0.0%, H_A: μ ≠ 0.0%. The test statistic (*z*-value) = (2.0 − 0.0) / (20.0 / 6) = 0.60. The significance level = 1.0 − 0.99 = 0.01, or 1%.

Since this is a two-tailed test with an alpha of 0.01, there is a 0.005 rejection region in both tails. Thus, we need to find the value 0.995 (1.0 − 0.005) in the table. The closest value is 0.9951, which corresponds to a critical *z*-value of 2.58. Since the test statistic is less than the critical value, we fail to reject H_0 and conclude that the stock's return equals 0.0%.

Cumulative *Z*-Table

$P(Z \leq z) = N(z)$ for $z \geq 0$

$P(Z \leq -z) = 1 - N(z)$



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

STUDENT'S *T*-DISTRIBUTION

Level of Significance for One-Tailed Test						
df	0.100	0.050	0.025	0.01	0.005	0.0005
Level of Significance for Two-Tailed Test						
df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.294
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

F-TABLE AT 5%

Critical values of the *F*-distribution at a 5% level of significance

Degrees of freedom for the numerator along top row

Degrees of freedom for the denominator along side row

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	6.01	2.94	2.90	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39

F-TABLE AT 2.5%

Critical values of the *F*-distribution at a 2.5% level of significance

Degrees of freedom for the numerator along top row

Degrees of freedom for the denominator along side row

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	648	799	864	900	922	937	948	957	963	969	977	985	993	997	1001	1006
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48

CHI-SQUARED TABLE

Values of χ^2 (Degrees of Freedom, Level of Significance)

Probability in Right Tail

Degrees of Freedom	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000157	0.000982	0.003932	0.0158	2.706	3.841	5.024	6.635	7.879
2	0.020100	0.050636	0.102586	0.2107	4.605	5.991	7.378	9.210	10.597
3	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170

INDEX

A

accuracy, 234
actions, 214
activation function, 233
AdaBoost, 231
adjusted R^2, 126
alternative hypothesis, 88
antithetic variate technique, 195
area under the ROC curve (AUC), 235
arithmetic means, 66
artificial neural networks (ANNs), 233
attenuation bias, 112
augmented Dickey-Fuller test, 174
autocorrelation function (ACF), 150
autocorrelation (ρ) cutoff, 156
autocovariance, 150
autocovariance function, 150
autoregressive moving average (ARMA) process, 158
autoregressive process, first-order, 154

B

bagging, 230
bag of words, 215
Bayes' rule, 5
Bernoulli distribution, 25
Bernoulli random variable, 25
best linear unbiased estimator (BLUE), 69, 144
beta distribution, 40
bias, 233
bias-variance tradeoff, 143, 211
binomial distribution, 25
binomial random variable, 25
boosting, 231
bootstrap aggregation, 230

bootstrapping method, 198
Box-Pierce (BP) statistic, 161

C

calendar effects, 170
central limit theorem (CLT), 70, 112
central moment, 15
central tendency, 65
centroids, 209
chi-squared distribution, 37
circular block bootstrap (CBB), 199
classification and regression trees (CARTs), 227
coefficient of determination, 111, 126
cokurtosis, 77
conditional distribution, 48
conditional expectations, 55
conditional heteroskedasticity, 138
conditionally independent event, 3
conditional probability, 1, 4
confidence interval, 30, 95
confidence interval of the slope coefficient, 115
confusion matrix, 234
consistent estimator, 112
continuous random variable, 12
continuous uniform distribution, 23
control variate technique, 196
Cook's measure, 144
correlation, 51, 77
coskewness, 53, 77
covariance, 50, 75
covariance stationary, 150
critical value, 89
cross-validation, 213
cumulative distribution function (CDF), 12, 24

D

data cleaning, 207
data generating process (DGP), 193

data preparation, 207
decision nodes, 227
decision rule, 90
decision trees, 227
deep reinforcement learning, 214
degrees of freedom, 128
dependent variable, 107
deterministic trends, 167
Dickey-Fuller distribution, 174
discrete probability function, 4
discrete random variable, 11
distributed lag, 157
dummy variables, 111, 170, 223

E

economic significance, 96
elastic net, 224
entropy, 228
error rate, 234
estimator, 67
Euclidean distance, 209
event, 2
event space, 2
excess kurtosis, 73
expected value, 13
explained sum of squares, 125, 246
explained variable, 107
explanatory variable, 107
exploitation, 214
exploration, 214
exponential distribution, 39

F

false negatives (FN), 234
false positives (FP), 234
F-distribution, 38
first differences, 174
first-order autoregressive [AR(1)] process, 154

first-order moving average [MA(1)] process, 155
F-statistic, 129
F-test, 129

G

Gaussian white noise, 152
generalization, 212
general linear process, 153
general-to-specific model, 143
Gini measure, 228
gradient descent algorithm, 233

H

hazard rate, 39
heteroskedasticity, 137
homoskedasticity, 137
h-step-ahead point forecast, 172
hyperparameter, 224
hypothesis, 87
hypothesis testing, 127

I

implied volatility, 181
independent and identically distributed (i.i.d.) random variables, 57, 70, 198
independent event, 2
independent variable, 107
independent white noise, 152
inertia, 210
information gain, 228
intercept, 110
interquartile range (IQR), 19, 75

J

Jarque-Bera (JB) test statistic, 182

joint probability, 2, 5

K

Kendall's τ (tau), 184
K-means algorithm, 209
K-nearest neighbors (KNN), 231
kurtosis, 16, 72

L

lag operator, 157
law of large numbers (LLN), 70
learning rate, 233
least absolute shrinkage and selection operator (LASSO), 224
leave-one-out cross validation, 213
lemmatization, 215
leptokurtic distribution, 73
linear time trend, 167
linear transformation, 19
Ljung-Box (LB) statistic, 161
logistic function, 225
logistic regression (logit) model, 225
log-likelihood function, 226
log-linear model, 168
lognormal distribution, 34
log-quadratic model, 169

M

machine learning, 205
Manhattan distance, 209
marginal distribution, 47
maximum likelihood method, 225
mean absolute forecast error, 226
mean squared forecast error (MSFE), 226
median, 19, 73
m-fold cross-validation, 143
mixture distributions, 40

model specification, 142
Monte Carlo method, 214
Monte Carlo simulation, 193
moving average [MA(1)] process, first-order, 155
moving average representation, 156
multicollinearity, 139, 146
multiple regression, 121
interpretation, 123
multiple testing, 102
multivariate random variables, 45
mutually exclusive event, 3

N

natural language processing (NLP), 215
negative skew, 71
neural networks, 214, 233
N-grams, 215
nonlinear time trend, 168
normal distribution, 29
normalization, 207
normal white noise, 152
null hypothesis, 88

O

omitted variable bias, 142
omitted variables, 112
one-hot encoding, 223
one-tailed test, 89
ordinary least squares (OLS), 109, 122
outliers, 71
overdifferencing, 174
overfitting, 211

P

partial autocorrelation, 160
partial autocorrelation function, 151

partial slope coefficients, 110, 122
Pearson's correlation, 184
perfect collinearity, 139
point estimate, 67
Poisson distribution, 27
polynomial time trend, 168
population mean, 65
positive definiteness, 187
positive skew, 71
post-pruning, 230
Power Law, 183
power of a test, 94
precision, 234
pre-pruning, 230
price relatives, 35
principal components analysis (PCA), 208
probability density function (PDF), 18
probability mass function (PMF), 12, 46
probability matrix, 46
pseudo-random number generators (PRNGs), 200
p-value, 97, 115

Q

quantile function, 18
quantiles, 74
quartiles, 74

R

random forests, 230
random variable, 1, 45
random walk, 173
recall, 234
receiver operating characteristic (ROC) curve, 235
regression analysis, 107, 121
regression coefficient, 127
regression function, 121
regularization, 224
reinforcement learning, 206

residual plots, 143
residual sum of squares, 125, 246
rewards, 214
ridge regression, 224
root node, 227

S

sample autocorrelation, 160
sample mean, 63
sample selection bias, 111
sampling error, 194
seasonal differencing, 171
seasonality, 170
serially uncorrelated, 152
shrinkage penalty, 224
silhouette coefficient, 210
simple return, 179
simultaneity bias, 112
skewness, 16, 71
slope coefficient, 109
Spearman's rank correlation, 184
standard deviation, 64
standard error of the regression (SER), 114, 125
standardization, 207
standard normal distribution, 31
states, 214
statistical significance, 96, 128
stemming, 215
stochastic trends, 167
stopwords, 215
structural changes, 199
Student's *t*-distribution, 36
supervised learning, 206
support vector machines (SVMs), 231
survivorship bias, 111

T

temporal difference learning method, 214

terminal nodes, 227
test set, 212
test statistic, 89
text mining, 215
time series, 149
tokenized, 215
total probability rule, 5
total sum of squares, 125, 246
training set, 212
true negatives (TN), 234
true positives (TP), 234
t-test, 98
two-tailed test, 89
Type I error, 93
Type II error, 93, 139

U

unbiased estimator, 69, 112
unconditional heteroskedasticity, 137
unconditional probability, 1
underfitting, 211
uniform distribution, 23
unit root, 173
unsupervised learning, 206

V

validation set, 212
variance, 15, 64, 67
variance inflation factor (VIF), 140
variance rate, 180
volatility, 180

W

white noise, 152
Gaussian, 152
independent, 152

normal, 152
Wold's theorem, 153

Y

Yule-Walker equation, 155

Z

z-distribution, 31
z-test, 99
z-value, 31, 32

Required Disclaimers:

CFA Institute does not endorse, promote, or warrant the accuracy or quality of the products or services offered by Kaplan. CFA Institute, CFA®, and Chartered Financial Analyst® are trademarks owned by CFA Institute.

Certified Financial Planner Board of Standards Inc. owns the certification marks CFP®, CERTIFIED FINANCIAL PLANNER™, and federally registered CFP (with flame design) in the U.S., which it awards to individuals who successfully complete initial and ongoing certification requirements. The College for Financial Planning®, a Kaplan company, does not certify individuals to use the CFP®, CERTIFIED FINANCIAL PLANNER™, and CFP (with flame design) certification marks. CFP® certification is granted only by Certified Financial Planner Board of Standards Inc. to those persons who, in addition to completing an educational requirement such as this CFP® Board-Registered Program, have met its ethics, experience, and examination requirements.

The College for Financial Planning®, a Kaplan company, is a review course provider for the CFP® Certification Examination administered by Certified Financial Planner Board of Standards Inc. CFP Board does not endorse any review course or receive financial remuneration from review course providers.

GARP® does not endorse, promote, review, or warrant the accuracy of the products or services offered by Kaplan of FRM® related information, nor does it endorse any pass rates claimed by the provider. Further, GARP® is not responsible for any fees or costs paid by the user to Kaplan, nor is GARP® responsible for any fees or costs of any person or entity providing any services to Kaplan. FRM®, GARP®, and Global Association of Risk Professionals™ are trademarks owned by the Global Association of Risk Professionals, Inc.

CAIAA does not endorse, promote, review or warrant the accuracy of the products or services offered by Kaplan, nor does it endorse any pass rates claimed by the provider. CAIAA is not responsible for any fees or costs paid by the user to Kaplan nor is CAIAA responsible for any fees or costs of any person or entity providing any services to Kaplan. CAIA®, CAIA Association®, Chartered Alternative Investment Analyst℠, and Chartered Alternative Investment Analyst Association® are service marks and trademarks owned by CHARTERED ALTERNATIVE INVESTMENT ANALYST ASSOCIATION, INC., a Massachusetts non-profit corporation with its principal place of business at Amherst, Massachusetts, and are used by permission.